On the Asymptotic Properties of Debiased Machine Learning Estimators *

Amilcar Velez
Department of Economics, Cornell University
amilcare@cornell.edu

This version: February 4, 2026.
Newest version here.

Abstract

This paper studies debiased machine learning (DML) under a novel asymptotic framework, providing insights that inform applied practice and explain simulation findings. DML is a two-step estimation method applicable to many econometric models where the parameter of interest depends on unknown nuisance functions. It uses K-fold sample splitting to estimate the nuisance functions and attains standard asymptotic properties under weaker conditions than classical semiparametric methods, accommodating flexible machine-learning estimators in the first step. Practitioners implementing DML confront two main decisions: whether to use DML1 or DML2 (the two variants of DML estimators), and how to choose K? Existing practice favors DML2 with large K based on simulation evidence, but these recommendations lack theoretical justification, as existing theory shows both variants are asymptotically equivalent for any fixed K. Under an asymptotic framework in which K grows with the sample size n, we demonstrate that DML2 offers theoretical advantanges over DML1 in terms of bias, mean squared error, and inference. We provide conditions under which increasing K reduces DML2's second-order asymptotic bias and MSE. These results support using DML2 with K as large as feasible, and in particular with $K = n$, for which we propose a computationally simple procedure.

KEYWORDS: Debiased machine learning, cross-fitting, second-order asymptotic approximation.

*I am deeply grateful to Ivan Canay, Federico Bugni, and Joel Horowitz for their guidance and support and for the extensive discussions that have helped shape the paper. I also want to acknowledge helpful conversations with Eric Auerbach, Federico Crippa, Jacob Dorn, Igal Hendel, Diego Huerta, Danil Fedchenko, Chuck Manski, Giorgio Primiceri, Sebastian Sardon, Chris Walker, and Thomas Wiemann as well as the good feedback provided by seminar participants at Duke, Rice, FGV-EPGE, UPenn, Rutgers, Michigan, Emory QTM, USC, Cornell, ESWC 2025, Cornell/PennState Conference on Econometrics and IO, Microeconometrics Class of 24/25, Georgetown U, XLIII Encuentro de Economistas del BCRP (Peru), LMU of Munich, University of Mannheim, and University of Bonn. Financial support from the Robert Eisner Memorial Fellowship and the Dissertation Year Fellowship at Northwestern University is gratefully acknowledged. Any and all errors are my own.

1 Introduction

Debiased machine learning (DML) has become a popular estimation method for econometric settings where the parameter of interest depends on unknown nuisance functions (Chernozhukov et al., 2018; Ahrens et al., 2025). DML attains standard asymptotic properties under milder conditions than classical semiparametric methods (Newey, 1994; Andrews, 1994; Newey and McFadden, 1994), accommodating machine-learning estimators for nuisance functions. In practice, two DML estimators can be used: DML1 and DML2. Both randomly divide the data into K equal-sized folds to estimate nuisance functions, but differ in how these estimates are combined. Practitioners implementing DML face two key questions: whether to use DML1 or DML2, and how to choose K? Existing recommendations favor DML2 with large K based on simulation evidence (Chernozhukov et al., 2018; Ahrens et al., 2024a, 2025). However, these recommendations lack theoretical justification, as existing asymptotic theory establishes that DML1 and DML2 have identical limiting distributions for any fixed K, providing no guidance for choosing between methods or selecting K. To address these questions, this paper studies the properties of DML1 and DML2 under a novel asymptotic framework in which K may grow with the sample size n. Under this framework, we demonstrate that DML2 weakly dominates DML1 in terms of bias, MSE, and inference. We also characterize when this dominance is strict and thereby explain simulation patterns that fixed-K theory cannot. We provide conditions under which increasing K reduces DML2's second-order asymptotic bias and MSE. These results support using DML2 with K as large as feasible, and in particular with $K = n$, for which we propose a computationally simple procedure.

DML is an estimation method applicable to econometric models in which the parameter of interest θ_0 is finite dimensional and satisfies a moment condition of the following form:

$$E[m(W, \theta_0, \eta_0)] = 0 \ , \tag{1.1}$$

where m is a known moment function, W is an observed random vector, and η_0 is an unknown nuisance function. Examples of a parameter θ_0 that can be identified by the moment condition (1.1) include several treatment effect parameters, such as the average treatment effect (ATE), average treatment effect on the treated in difference-in-differences designs (ATT-DID), local average treatment effect (LATE), weighted average treatment effects (w-ATE), average treatment effect on the treated (ATT), treatment effect coefficient in the partial linear model (PLM), among others, all of which have been studied in the literature on semiparametric models (e.g., Robinson (1988), Robins et al. (1994), Hahn (1998), Hirano et al. (2003), Frölich (2007), Farrell (2015), Chernozhukov et al. (2017), Sant'Anna and Zhao

(2020), and Chang (2020)). In all these examples, the moment function m is linear in the parameter θ_0, and the nuisance function η_0 is based on conditional expectations, such as the propensity score. This paper considers a setup that includes all these examples.

DML relies on two ingredients to attain standard asymptotic properties (e.g., asymptotic normality with parametric rates) for the DML estimator of θ_0. The first one is the *Neyman orthogonality*, a necessary condition on the moment function m to guarantee that the estimation of θ_0 is as accurate as if the true η_0 had been used; see Remarks 3.1 and 3.2. The second ingredient is *cross-fitting*, a form of sample splitting used in the nuisance function estimation, that complements the orthogonality condition to accommodate for a larger class of flexible nuisance functions estimators, including machine-learning methods.

Two versions of the DML estimator for θ_0 were proposed by Chernozhukov et al. (2018), namely DML1 and DML2. Both versions randomly divide the data into K equal-sized folds, denoted as $\mathcal{I}_k$ for $k = 1, \ldots, K$. For each fold $\mathcal{I}_k$, an estimator $\hat{\eta}_k$ of η_0 is constructed using all the data except the data in fold $\mathcal{I}_k$. Then, DML1 first calculates preliminary estimators $\tilde{\theta}_k$ by solving the moment condition (1.1) within each fold $\mathcal{I}_k$ using the estimator $\hat{\eta}_k$. It then combines the information across the folds by averaging the $\tilde{\theta}_k$'s to obtain the proposed estimator for θ_0. In contrast, DML2 first combines the information across the folds by averaging moment conditions based on (1.1), where each fold uses estimates $\hat{\eta}_k$, and then θ_0 is estimated as the solution in θ of the average of moment conditions, $K^{-1}\sum_{k=1}^{K}\left((n/K)^{-1}\sum_{i\in\mathcal{I}_k} m(W_i, \theta, \hat{\eta}_k)\right) = 0$.

Practitioners implementing DML confront two main decisions: whether to use DML1 or DML2, and how to choose K? The literature already recommends DML2 over DML1, and suggests using a large K for DML2 based on simulation evidence (Chernozhukov et al., 2018; Ahrens et al., 2024a, 2025). However, these recommendations lack theoretical justification, since existing asymptotic theory predicts that DML1 and DML2 have the same limiting distribution when the number of folds K remains fixed as the sample size n tends to infinity.

To address these questions, this paper studies the properties of DML1 and DML2 under a novel asymptotic framework in which K may grow with the sample size n. This asymptotic framework captures finite-sample situations in which practitioners desire to use a large K to improve the precision of the estimators $\hat{\eta}_k$'s, which use a fraction $(K-1)/K$ of the data. This approach follows a tradition in econometrics of using refined asymptotic approximations to study finite-sample behavior, as in Cattaneo and Jansson (2018), Bugni and Canay (2021), and Cai (2022). Under this framework, we can distinguish between DML1 and DML2 and characterize how K affects their performance, thereby explaining simulation patterns that fixed-K asymptotic theory cannot and providing formal guidance for implementation.

We make three contributions. First, we provide an asymptotic result that explains the

discrepancy found in simulations between DML1 and DML2. Formally, we demonstrate that DML2 weakly dominates DML1 in terms of bias, MSE, and inference. We also characterize when this dominance is strict, and thereby explain simulation patterns that fixed-K asymptotic theory cannot.

Second, we show that the existing estimation and inference results for DML2—based on fixed-K asymptotic theory—continue to apply for any $K \in \{2, \dots, n\}$. Formally, we provide conditions under which the finite-sample distribution of DML2 is approximated by the same limiting distribution uniformly in K. This implies that we can use DML2 with $K = n$, i.e., the leave-one-out estimator, which ensures replicability.

Third, we derive a second-order asymptotic approximation for scalar DML2 estimators that we use to explain observed patterns in DML2's finite-sample bias and MSE. Under the conditions we provide, we conclude that increasing K decreases the second-order asymptotic bias and MSE, implying that an optimal choice for DML2 is $K = n$. In particular, commonly recommended choices such as $K = 5$ or $K = 10$ are suboptimal. Furthermore, we use our approximation to quantify the relatively efficiency loss from suboptimal choices.

Finally, we use our theoretical results to provide three recommendations for the implementation of DML. First, practitioners should prefer DML2 over DML1, because DML2 offers theoretical advantages over DML1, and DML2 is robust to the choice of K for a large class of first-step estimators. Second, practitioners should use DML2 with $K = n$. This choice of K is optimal and ensures replicability of the estimator. We propose a computationally simple procedure to implement DML2 with $K = n$ in Section 5.2. Lastly, if practitioners must choose a small value of K such as $K = 5$ or $K = 10$, they should prefer $K = 10$ since this choice guarantees substantially lower efficiency losses than $K = 5$.

Related Literature: This paper contributes to the growing DML literature, where estimators have been developed for semiparametric problems without requiring strong conditions on nuisance estimators (e.g., without Donsker class assumptions). Examples include Chernozhukov et al. (2017), Chernozhukov et al. (2018), Chernozhukov et al. (2022a), Chernozhukov et al. (2022b,c), Semenova and Chernozhukov (2021), Semenova (2023a,b), Escanciano and Terschuur (2023), Rafi (2023), Cheng et al. (2023), Ji et al. (2023), Noack et al. (2024), Fava (2024), Kennedy et al. (2024), and Jin and Syrgkanis (2024). Most of these papers use DML2, with exceptions such as Chernozhukov et al. (2017), Ji et al. (2023), and Cheng et al. (2023), which use DML1.[1] With the exception of Kennedy et al. (2024) and Jin and Syrgkanis (2024), these papers derive first-order asymptotic theory assuming K remains fixed as $n \to \infty$. Kennedy et al. (2024) and Jin and Syrgkanis (2024) use a

[1] In many of these papers, such as Rafi (2023) and Semenova (2023a), DML1 and DML2 are numerically equivalent; see Remark 2.1.

structure-agnostic framework to establish optimality of DML estimators. In contrast, we study DML1 and DML2 properties when $K \to \infty$ as $n \to \infty$, demonstrating that DML2 offers theoretical advantages over DML1 and characterizing conditions under which DML2 with $K = n$ (the leave-one-out estimator) is optimal in terms of second-order asymptotic bias and MSE. To the best of our knowledge, this literature does not provide theoretical guidance for selecting K, an issue we address here.

This paper also contributes to the literature on double-robust estimators, including Robins et al. (1994), Robins and Rotnitzky (1995), Scharfstein et al. (1999), Farrell (2015), Sant'Anna and Zhao (2020), Chang (2020), Callaway and Sant'Anna (2021), Rothe and Firpo (2019), and Singh and Sun (2024). With the exception of Rothe and Firpo (2019), these papers study first-order asymptotic theory for estimators that remain consistent even when some components of η_0 are misspecified. Rothe and Firpo (2019) studies higher-order properties of double-robust estimators in a missing-data setting where η_0 is estimated via a leave-one-out approach. Our results complement that work in several ways. First, the DML versions of double-robust estimators accommodate flexible estimation of η_0 components. Second, we present second-order properties of DML2 estimators more generally. Third, we show that among DML2 estimators, the leave-one-out estimator is optimal in terms of bias and MSE under certain conditions. Notably, this optimal estimator coincides with the estimator studied in Rothe and Firpo (2019).

More broadly, this paper contributes to the literature on semiparametric models, which has a long tradition in econometrics and statistics (e.g., Bickel (1982), Robinson (1988), Newey (1990), Andrews (1994), Newey and McFadden (1994), Newey (1994), Linton (1995), and Bickel and Ritov (2003)). Many papers in this literature provide conditions for studying estimators based on a plug-in approach (i.e., the same data are used to estimate η_0 and θ_0). In contrast, we provide conditions for studying the second-order properties of DML2, which uses sample splitting.

Outline: Section 2 presents notation and summarizes existing results. Section 3 presents the limiting distributions of DML1 and DML2 for large K values. Section 4 derives the second-order asymptotic approximation for DML2 allowing for large K values. Section 5 presents recommendations based on our theoretical results. Section 6 shows simulation evidence that motivate the problem. Section 7 concludes. Appendix contains the proofs of main results, auxiliary results, and well-know examples in the literature. Supplemental Appendix contain additional simulations and proofs.

Notation: We use: $[K] = \{1, \ldots, K\}$, $[n] = \{1, \ldots, n\}$, $[p] = \{1, \ldots, p\}$, $||\cdot||$ denotes the euclidean distance for vectors and the L_2-operator norm for matrices, and $||\cdot||_\infty$ denotes the element-wise supremum norm for vectors and matrices.

2 Setup and Previous Results

The parameter of interest is $\theta_0 \in \Theta \subseteq \mathbf{R}^d$ and satisfies the following moment condition:

$$E[m(W, \theta_0, \eta_0(X))] = 0_{d\times 1} \ , \tag{2.1}$$

where $m : \mathcal{W} \times \Theta \times \mathcal{E} \to \mathbf{R}^d$ is a known moment function and $(W, X) \in \mathcal{W} \times \mathcal{X} \subseteq \mathbf{R}^{d_w+d_x}$ is a random vector with distribution F_0. The nuisance function $\eta_0 : \mathcal{X} \subseteq \mathbf{R}^{d_x} \to \mathcal{E} \subseteq \mathbf{R}^p$ is an unknown function of the covariates X.

This paper considers moment functions m that are linear in the parameter of interest:

$$m(W, \theta, \eta) = \psi^b(W, \eta) - \psi^a(W, \eta)\theta \ , \tag{2.2}$$

where ψ^b and ψ^a are functions that satisfy conditions specified in Assumption 3.1, which includes the identification condition, $E[\psi^a(W, \eta_0(X))] \in \mathbf{R}^{d\times d}$ is invertible, and guarantees a Neyman orthogonality condition,

$$E\left[\partial_\eta m(W, \theta_0, \eta_0(X)) \mid X\right] = 0 \ , \quad a.e.$$

where $\partial_\eta m$ denotes the matrix of partial derivatives m with respect to the values of η and $\partial_\eta m(W, \theta_0, \eta_0(X))$ is the $\partial_\eta m$ evaluated at $\eta = \eta_0(X)$.

A wide range of parameters of interest can be identified through moment conditions such as (2.1) using a moment function like (2.2). Examples of θ_0 include the average treatment effect (Example C.1), the average treatment effect on the treated in difference-in-differences designs (Example C.2), and the local average treatment effect (Example C.3), among others. All these examples are presented in Appendix C and additional examples appear in Ahrens et al. (2025).

Consider the goal of estimating θ_0 using a random sample $\{(W_i, X_i) : 1 \le i \le n\}$ drawn from the distribution F_0. The parameter θ_0 based on (2.1) and (2.2) can be identified as follows,

$$\theta_0 = E[\psi^a(W, \eta_0(X))]^{-1} \ E[\psi^b(W, \eta_0(X))] \ . \tag{2.3}$$

Accordingly, an ideal estimator for θ_0 is defined by replacing the expected values in (2.3) with sample analogs. That is,

$$\hat{\theta}_n^* = \left(n^{-1}\sum_{i=1}^n \psi^a(W_i, \eta_i)\right)^{-1} \left(n^{-1}\sum_{i=1}^n \psi^b(W_i, \eta_i)\right) \ , \tag{2.4}$$

where $\eta_i = \eta_0(X_i)$ denotes the value of the nuisance function η_0 evaluated at the covariate X_i for observation i. However, the values of the η_i's are unknown. As a result, the oracle estimator $\hat{\theta}_n^*$ is infeasible. For this reason, it is common to calculate first estimates $\hat{\eta}_i$ of η_i that can be used later to compute an estimator of θ_0. Remark 2.4 discusses the *plug-in* approach used in classical semiparametric methods. In what follows, we formally explain how DML estimates η_0 and θ_0.

2.1 First-Step DML: Nuisance Function Estimation

DML proposes to calculate the estimates $\hat{\eta}_i$ of η_i using a *cross-fitting* procedure, which is a form of sample-splitting. This procedure has two steps and implicitly assumes that n can be divided by K:[2]

1. *Sample splitting*: Randomly split the indices into K equal-sized folds $\mathcal{I}_k$, i.e., $\cup_{k=1}^K \mathcal{I}_k = [n]$. The number of observations in fold $\mathcal{I}_k$ is denoted by $n_k = n/K$.

2. *Nuisance Function Estimates*: For each fold $\mathcal{I}_k$, the estimates $\hat{\eta}_i$ of η_i are defined by

$$\hat{\eta}_i = \hat{\eta}_k(X_i) \ , \quad \forall i \in \mathcal{I}_k \ , \tag{2.5}$$

   where $\hat{\eta}_k(\cdot)$ is an estimator of the nuisance function $\eta_0(\cdot)$ using $\{W_i : i \notin \mathcal{I}_k\}$, which is all the data except the ones with indices on the fold $\mathcal{I}_k$. All the estimates $\hat{\eta}_i$ are calculated by repeating the process for all the $k \in [K]$.

Both DML estimators use the same estimates $\hat{\eta}_i$, but they differ in how they combine information across the different folds defined above. We explain this next.

2.2 Second-Step DML: Parameter-of-Interest Estimation

Definition 2.1 (DML1)**.** The DML1 estimator first calculates preliminary estimators $\tilde{\theta}_k$ by solving the moment condition (2.1) within each fold $\mathcal{I}_k$ using the estimates $\hat{\eta}_i$,

$$\tilde{\theta}_k \quad \text{solve} \quad n_k^{-1} \sum_{i \in \mathcal{I}_k} m(W_i, \theta, \hat{\eta}_i) = 0 \ ,$$

[2]When n is not divisible by K, the number of observations in some folds will be $\lfloor n/K \rfloor$ while in others $\lfloor n/K \rfloor + 1$, where $\lfloor n/K \rfloor$ is the greatest integer less than or equal to n/K.

it then combines the information across the folds by averaging the $\tilde{\theta}_k$'s to obtain the proposed estimator for θ_0,

$$\hat{\theta}^{(1)}_{n,K} = K^{-1} \sum_{k=1}^{K} \tilde{\theta}_k \ . \tag{2.6}$$

Explicit expressions for $\tilde{\theta}_k$ can be obtained since the moment function m is as in (2.2),

$$\tilde{\theta}_k = \left(\frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \psi^a(W_i, \hat{\eta}_i) \right)^{-1} \left(\frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \psi^b(W_i, \hat{\eta}_i) \right) \ , \quad \forall \ k \in [K] \ .$$

Note that $\tilde{\theta}_k$ is similar to (2.4) but using only observations in the fold $\mathcal{I}_k$ and the estimates $\hat{\eta}_i$ instead of η_i.

Definition 2.2 (DML2)**.** The DML2 estimator first combines the information across the folds $\mathcal{I}_k$ by averaging the sample analog of moment conditions like (2.1) using the estimates $\hat{\eta}_i$, and then estimates θ_0 by solving the average of moment conditions,

$$\hat{\theta}^{(2)}_{n,K} \quad \text{solve} \quad \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{i \in \mathcal{I}_k} m(W_i, \theta, \hat{\eta}_i) \right) = 0 \ .$$

An explicit expression for $\hat{\theta}^{(2)}_{n,K_n}$ is obtained by using that the moment function m is as in (2.2),

$$\hat{\theta}^{(2)}_{n,K} = \left(\frac{1}{n} \sum_{i=1}^{n} \psi^a(W_i, \hat{\eta}_i) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \psi^b(W_i, \hat{\eta}_i) \right) \ . \tag{2.7}$$

Note that $\hat{\theta}^{(2)}_{n,K}$ is similar to (2.4) but using the estimates $\hat{\eta}_i$ instead of η_i.

Remark 2.1. The DML1 and DML2 estimators can be equal under certain conditions. If $\psi^a(W_i, \hat{\eta}_i)$ has zero variance (e.g., ψ^a is a constant ψ^a_0 as in Example C.1) and the K-fold partition $\{I_k : 1 \leq k \leq K\}$ divides the data into exactly K subsets with equal size, then both DML1 and DML2 estimators defined in (2.6) and (2.7) are equal. In particular,

$$\begin{aligned}
\hat{\theta}^{(1)}_{n,K} &= K^{-1} \sum_{k=1}^{K} (\psi^a_0)^{-1} \left(n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^b(W_i, \hat{\eta}_i) \right) \\
&= (\psi^a_0)^{-1} \left(n^{-1} \sum_{i=1}^{n} \psi^b(W_i, \hat{\eta}_i) \right) \\
&= \hat{\theta}^{(2)}_{n,K} \ .
\end{aligned}$$

Therefore, the DML1 and DML2 estimators for the ATE (Example C.1) are numerically the same when the data are divided in exactly K folds. In contrast, if $\psi^a(W_i, \hat{\eta}_i)$ has positive variance, then $\hat{\theta}^{(1)}_{n,K} \neq \hat{\theta}^{(2)}_{n,K}$ in general. This occurs in all the other examples. □

Remark 2.2. It has been recommended to use DML2 based on simulation evidence. More concretely, Remark 3.1 in Chernozhukov et al. (2018) state that for some moment functions there is no difference between DML1 and DML2, but for some other moment functions DML2 is better behaved than DML1. The theoretical reasons explaining this difference are unknown. Section 3 will provide an explanation. □

Remark 2.3. Simulation evidence shows that increasing the number of folds K reduces DML2's finite-sample bias and MSE (Ahrens et al., 2024a,b; Chernozhukov et al., 2018)). An intuitive explanation is that more accurate first-step estimators should imply better estimators for the parameter of interest. One way to obtain more data in the first-step is to increase K, since the nuisance function estimators use 50%, 80%, and 90% of the data when K is 2, 5, and 10, respectively. However, it is theoretically unknown if the improvement in the accuracy of the estimation of η_0 translated into more precise estimates for θ_0. Section 4 will provide conditions that formalize this intuition. □

Remark 2.4. An estimator $\hat{\eta}$ of η_0 can be obtained by using all the data, and then an estimator of θ_0 can be defined by replacing η_i by the estimates $\hat{\eta}_i$ in (2.4), where $\hat{\eta}_i = \hat{\eta}(X_i)$. An estimator of θ_0 based on this approach is known as the plug-in estimator, and the conditions under which it has standard properties (e.g., asymptotic normality and parametric convergence rates) have been studied in the literature on semiparametric models (e.g., Andrews (1994), Newey (1994), Newey and McFadden (1994)). However, this approach is sensitive to the "own observation" bias, which arises when the same data are used to estimate both η_0 and θ_0 (Newey and Robins (2018)). Stronger conditions are often required on the class of nuisance functions to attenuate the own observation bias on the analysis of the plug-in estimator. In contrast, DML—the approach considered in this paper to construct estimators—removes this bias by relying on cross-fitting, which is a form of sample splitting, that allocates one part of the data to estimate the nuisance function and other to estimate the parameter of interest.

□

2.3 Previous Results

Under some conditions (including K fixed as $n \to \infty$), Chernozhukov et al. (2018) showed that both DML estimators $\hat{\theta}^{(1)}_{n,K}$ and $\hat{\theta}^{(2)}_{n,K}$ have the same asymptotic distribution,

$$\sqrt{n}\left(\hat{\theta}^{(j)}_{n,K} - \theta_0\right) \xrightarrow{d} N(0, \Sigma) \ , \tag{2.8}$$

where the variance of the asymptotic distribution is given by

$$\Sigma = E\left[\psi^a(W, \eta_0(X))\right]^{-1} E\left[m(W, \theta_0, \eta_0(X))m(W, \theta_0, \eta_0(X))^\intercal\right] E\left[\psi^a(W, \eta_0(X))\right] \ , \tag{2.9}$$

which only depends on the moment function m, the true nuisance function η_0, and the data distribution F_0. This result implies that the existing theoretical framework cannot distinguish between estimators based on DML1 and DML2, as discussed in the introduction. Moreover, this asymptotic theory provides no direct guidance to select K for DML2.

The proof of (2.8) relies on a first-order equivalent condition, which is the central idea in DML. More concretely, both DML estimators $\hat{\theta}^{(1)}_{n,K}$ and $\hat{\theta}^{(2)}_{n,K}$ are first-order equivalent to the oracle estimator $\hat{\theta}^*_n$,

$$\sqrt{n}\left(\hat{\theta}^{(j)}_{n,K} - \hat{\theta}^*_n\right) \xrightarrow{p} 0 \ , \quad j = 1, 2 \ . \tag{2.10}$$

This result is the central idea since it implies that the estimation of θ_0 using DML is as accurate as if the true η_0 had been used.

Although the existing asymptotic theory shows that DML1 and DML2 are asymptotically equivalent, it has been conjectured that (1) DML2 can outperform DML1, as suggested by simulation evidence on their relative performance, and (2) increasing the number of folds K improves DML2 in terms of bias and mean-squared error (MSE). To investigate these conjectures, we consider an asymptotic framework in which K depends on the sample size n, i.e., $K = K_n$, allowing $K_n \to \infty$ as $n \to \infty$. We show that the accuracy (bias and MSE) and inference of DML1 are sensitive to the choice of K, whereas DML2 is not, implying that DML2 offers theoretical advantage over DML1 in terms of bias, MSE, and inference. Moreover, we provide conditions under which increasing K reduces the magnitude of the second-order asymptotic bias and the second-order asymptotic MSE of DML2, leading to the practical recommendation of $K_n = n$ as an asymptotically optimal choice. Finally, Section 5.2 proposes a computationally simple procedure for implementing DML2 with $K_n = n$.

3 Asymptotic Theory for DML1 and DML2 when K increases

We show that DML2 offers theoretical advantages over DML1 when $K = K_n$ increases with the sample size n (Theorem 3.1). We also prove that DML2 is robust to the choice of $K = K_n$ for a large class of first-step estimators (Theorem 3.2). Finally, Section 3.1 explains why and when DML1 is sensitive to the choice of K. In what follows, we first present and discuss the assumptions, and we then establish the results.

Let $G = E[\psi^a(W, \eta_0(X))] \in \mathbf{R}^{d\times d}$ and $\Omega = E[m(W, \theta_0, \eta_0(X))\, m(W, \theta_0, \eta_0(X))^\intercal] \in \mathbf{R}^{d\times d}$. We write $\psi^a(W, \eta) = [\psi^a_{t,s}(W, \eta)]_{t,s}$ and $m(W, \theta, \eta) = [m_t(W, \theta, \eta)]_t$. Recall $\eta_i = \eta_0(X_i)$. Let c_G, c_1, and c_2 be positive constants. The next assumption restricts the class of econometric models through conditions on the moment function m and function ψ^a.

Assumption 3.1. *$m(W, \theta, \eta)$ and $\psi^a(W, \eta)$ are twice continuously differentiable with respect to $\eta \in \mathcal{E} \subseteq \mathbf{R}^p$ and satisfy*

(a) G and Ω are non-singular and $||G^{-1}|| < c_G$.

(b) $E[|m_t(W_i, \theta_0, \eta_i)|^4] < c_1$ and $E[|\psi^a_{t,s}(W_i, \eta_i)|^4] < c_1$ for all $t, s \in [p]$.

(c) $E[\partial_\eta m_t(W_i, \theta_0, \eta_i)|X_i] = 0$, $||E[(\partial_\eta m_t(W_i, \theta_0, \eta_i))(\partial_\eta m_t(W_i, \theta_0, \eta_i))^\intercal|X_i]||_\infty < c_2$, and $\sup_{\eta\in\mathcal{E}} ||\partial^2_\eta m_t(W_i, \theta_0, \eta)||_\infty \leq c_2$ for $t \in [p]$.

(d) $E[\partial_\eta \psi^a_{t,s}(W_i, \eta_i) \mid X_i] = 0$, $||E[(\partial_\eta \psi^a_{t,s}(W_i, \eta_i))(\partial_\eta \psi^a_{t,s}(W_i, \eta_i))^\intercal \mid X_i]|| < c_2$, and $\sup_{\eta\in\mathcal{E}} ||\partial^2_\eta \psi^a_{t,s}(W_i, \eta)||_\infty \leq c_2$ for all $t, s \in [p]$.

Parts (a) and (b) of Assumption 3.1 are standard conditions that guarantee identification of the parameter of interest and stochastic expansions for the oracle estimator, similar to Newey and Smith (2004, Assumptions 2 and 3). Part (c) of Assumption 3.1 presents a Neyman-orthogonality condition, $E[\partial_\eta m_t(W_i, \theta_0, \eta_i)|X_i] = 0$, involving standard partial derivatives as in Belloni et al. (2017) and Farrell et al. (2025) rather than functional derivatives as in Chernozhukov et al. (2018). This type of condition is necessary to guarantee that feasible estimators are as accurate as if the true values of the nuisance functions were used; see Remark 3.1 for an explanation. It is possible to transform a moment function into one that satisfies a Neyman-orthogonality condition under certain conditions; see Remark 3.2 for comments on existing methods. Part (c) of Assumption 3.1 also includes standard conditions ensuring that nonlinear effects of first-step estimation error are negligible when the nuisance estimators are sufficiently accurate (e.g., when their L_2-convergence rates are faster than $n^{-1/4}$). Finally, part (d) of Assumption 3.1 implies that we can construct a DML

estimator for each component of $G = E[\psi^a(W_i, \eta_i)]$. It holds automatically when ψ^a does not depend on η. Further discussion of part (d) of Assumption 3.1 appears in Section 3.1.

Assumption 3.1 holds in many common econometric models studied in the literature; see Appendix C and Ahrens et al. (2025) for several examples. However, it excludes settings in which the moment function is not differentiable in η; see Remark 3.3 for examples of such non-smooth models where DML estimators have been proposed.

We next present conditions on the class of first-step estimators.

Assumption 3.2. *For any sequence $K_n \leq n$,*

$$\sup_{1 \leq k \leq K_n} E[||\hat{\eta}_k(X) - \eta_0(X)||^2]^{1/2} = o(n^{-1/4}) \ .$$

Assumption 3.2 holds for several firs-step estimators considered in the DML literature. For instance, it holds for deep neural networks as in Farrell et al. (2021) and Schmidt-Hieber (2020). Under certain conditions, it holds for LASSO and related penalized estimators of linear models (Tibshirani (1996), Van de Geer (2008), Belloni et al. (2011)). Kernel estimators and series estimators (Newey (1997), Belloni et al. (2015), Chen (2007)) can verify this assumption after appropriate trimming to ensure bounded inverse density weights for kernel estimators, or well-conditioned Gram matrices for series estimators. It is unknown if this assumption holds for random forest; see Chi et al. (2022).

As is common in DML, a Neyman orthogonality condition on the moment function (Assumption 3.1) and sufficiently accurate first-step estimators (Assumption 3.2) are enough to derive the limiting distribution of DML estimators. The next theorem presents the limiting distributions of DML1 and DML2 under our new asymptotic framework, in which $K = K_n \to \infty$ as $n \to \infty$.

Theorem 3.1. *Let Assumptions 3.1 and 3.2 hold and let K_n be such that $K_n \leq n$ and $K_n/\sqrt{n} \to c \in [0, \infty)$ as $n \to \infty$. Then,*

$$\sqrt{n}\left(\hat{\theta}^{(1)}_{n,K_n} - \theta_0\right) \xrightarrow{d} N(c\Lambda, \Sigma)$$

and

$$\sqrt{n}\left(\hat{\theta}^{(2)}_{n,K_n} - \theta_0\right) \xrightarrow{d} N(0, \Sigma) \ ,$$

where $\hat{\theta}^{(1)}_{n,K_n}$, $\hat{\theta}^{(2)}_{n,K_n}$, and Σ are defined in (2.6), (2.7), and (2.9), respectively, and

$$\Lambda = -G^{-1} E\left[\psi^a(W, \eta_0(X)) \ G^{-1} \ m(W, \theta_0, \eta_0(X))\right] \ . \tag{3.1}$$

Theorem 3.1 provides an asymptotic result that explains the discrepancy found in sim-

ulations between DML1 and DML2. As we mentioned in Remark 2.2, DML2 has been recommended over DML1 based on simulation evidence. This theorem now provides the theoretical explanation. It shows that DML2 is asymptotically better than DML1 in terms of bias and MSE when $c > 0$ and $\Lambda \neq 0$, otherwise both share the same limiting distribution. Our explanation through Λ emerges under the proposed asymptotic framework, providing insights not captured by the existing asymptotic theory or simulation-based evidence.

The distinction between DML1 and DML2 relies on the *discrepancy measure* Λ that depends on the econometric model (m, θ_0, and η_0) and not on the first-step estimator $\hat{\eta}$. Therefore, the relative performance between DML1 and DML2 can be obtained by calculating Λ without using finite data. In particular, for several econometric models—such as ATE, ATT-DID, ATT, and PLM—$\Lambda = 0$, but for others like LATE and w-ATE, it is typically nonzero. Note that this discrepancy measure can be computed also for econometric models where the moment function is not differentiable in η. We argue that if for those models $\Lambda \neq 0$, then DML2 should be preferred over DML1 even if those models are outside the scope of our setup. We postpone our explanation to Section 3.1.

When $\Lambda \neq 0$, DML1 becomes increasingly sensitive to large K_n values regarding bias, MSE, and coverage probability of its associated confidence interval. In contrast, DML2 remains unaffected by the choice of K_n. By setting $c = K_n/\sqrt{n}$ and using the limiting distribution of DML1 in Theorem 3.1, we can approximate the finite sample distribution of DML1, $\sqrt{n}\left(\hat{\theta}^{(1)}_{n,K_n} - \theta_0\right)$, by $N((K_n/\sqrt{n})\Lambda, \Sigma)$ which is sensitive to the choice of K_n when n is small and $\Lambda \neq 0$. Intuitively, this suggest that the distribution of DML1 is not centered at the origin and the gap is increasing on K_n. This implies that the standard recommended DML1 confidence interval is not valid when $\Lambda \neq 0$ and, furthermore, its coverage decrease as K_n increases, explaining their simulation performance. We formalize this intuition in the next corollary.

Corollary 3.1. *Let $CI^{(1)}_\alpha$ be the standard recommended DML1 confidence interval for $\theta_{0,t}$ (t-th component of θ_0),*

$$CI^{(1)}_\alpha = [\hat{\theta}^{(1)}_{n,K_n} - z_{1-\alpha/2}\tfrac{\hat{s}_{t,n}}{\sqrt{n}}, \hat{\theta}^{(1)}_{n,K_n} + z_{1-\alpha/2}\tfrac{\hat{s}_{t,n}}{\sqrt{n}}] \ ,$$

where $\hat{s}^2_{t,n}$ is a consistent estimator for $\Sigma_{t,t}$ (t-th diagonal term of Σ). Under the conditions of Theorem 3.1, we then have

$$P\left(\theta_{0,t} \in CI^{(1)}_\alpha\right) = \mathcal{P}(K_n) + o(1) \ ,$$

where

$$\mathcal{P}(K_n) = \Phi\left(z_{1-\alpha/2} + \frac{K_n}{\sqrt{n}}\frac{\Lambda_t}{\sqrt{\Sigma_{t,t}}}\right) + \Phi\left(z_{1-\alpha/2} - \frac{K_n}{\sqrt{n}}\frac{\Lambda_t}{\sqrt{\Sigma_{t,t}}}\right) - 1 \ .$$

Here, $\Lambda = [\Lambda_t]_t$ *and* $\Sigma = [\Sigma_{t,s}]_{t,s}$*. Furthermore,* $\mathcal{P}(K_n)$ *is a decreasing function on* K_n *if and only if* $\Lambda_t \neq 0$*. In particular,* $\mathcal{P}(0) = 1 - \alpha > \mathcal{P}(K_n)$ *if* $K_n > 1$ *and* $\Lambda_t \neq 0$*.*

Theorem 3.1 shows that the finite-sample distribution of DML2 can be approximated by the limiting distribution implied by the existing fixed-K asymptotic theory, regardless of the choice of $K = K_n$, provided that $K_n = O(\sqrt{n})$. In particular, this suggests that DML2 inference is robust to the choice of K, which we formalize in the next corollary.

Corollary 3.2. *Under the conditions of Theorem 3.1, we then have*

$$P\left(\theta_{0,t} \in CI_\alpha^{(2)}\right) = 1 - \alpha + o(1) \ ,$$

where $CI_\alpha^{(2)}$ *is the standard recommended DML2 confidence interval for* $\theta_{0,t}$ *(*t*-th component of* θ_0*),*

$$CI_\alpha^{(1)} = [\hat{\theta}_{n,K_n}^{(1)} - z_{1-\alpha/2}\frac{\hat{s}_{t,n}}{\sqrt{n}}, \hat{\theta}_{n,K_n}^{(1)} + z_{1-\alpha/2}\frac{\hat{s}_{t,n}}{\sqrt{n}}] \ ,$$

where $\hat{s}_{t,n}^2$ *is a consistent estimator for* $\Sigma_{t,t}$ *(*t*-th diagonal term of* Σ*). Furthermore, the standard DML2 estimator* $\hat{s}_{t,n}^2$ *for* $\Sigma_{t,t}$ *defined in Chernozhukov et al. (2018, Theorem 3.2) is consistent under the conditions of Theorem 3.1.*

Theorem 3.1 has shown that DML2 asymptotically dominates DML1 and is robust to the choice of K, provided that $K = K_n = O(\sqrt{n})$. The next assumption is proposed to extend the robustness of DML2 to the choice of K from $K_n = O(\sqrt{n})$ to $K_n = O(n)$. Let $\hat{\eta}_k^\ell(\cdot)$ be the same estimator as $\hat{\eta}_k(\cdot)$ except in the use of observation ℓ: $\hat{\eta}_k(\cdot)$ uses (W_ℓ, X_ℓ), whereas $\hat{\eta}_k^\ell(\cdot)$ uses $(\widetilde{W}_\ell, \widetilde{X}_\ell)$, where the random vector $(\widetilde{W}_\ell, \widetilde{X}_\ell)$ is draw from F_0 and independent of the data (i.e., $(\widetilde{W}_\ell, \widetilde{X}_\ell)$ and (W_ℓ, X_ℓ) are i.i.d.).

Assumption 3.3. *For any sequence* $K_n \leq n$*,*

$$\sup_{1\leq k\leq K_n} \max_{\ell \notin \mathcal{I}_k} E[||\hat{\eta}_k(X) - \hat{\eta}_k^\ell(X)||^2]^{1/2} = o(n^{-1/2})$$

Assumption 3.3 is an *algorithm stability* condition similar to the one in Chen et al. (2022, Corollary 4). This condition measure the stability of first-step estimators to replacement of exactly one observation in L_2-norm. The study of this type of conditions has received considerable attention in the statistical machine learning and generalization theory literature; see Bousquet and Elisseeff (2002) and Hardt et al. (2016). This condition can be verified for kernel and series estimators after appropriate trimming to ensure bounded inverse density

weights for kernel estimators, or well-conditioned Gram matrices for series estimators. It is unknown if this condition can be verified for deep neural networks or other machine learning methods, which is an interesting research direction outside the scope of this paper.

The next theorem guarantees that DML2 is robust to the choice of K, provided that the first-step estimators are stable to replacing a single observation with another i.i.d. draw.

Theorem 3.2. *Let Assumptions 3.1, 3.2, and 3.3 hold and let K_n be such that $K_n \leq n$. Then,*

$$\sqrt{n}\left(\hat{\theta}^{(2)}_{n,K_n} - \theta_0\right) \xrightarrow{d} N(0, \Sigma) \ ,$$

where $\hat{\theta}^{(2)}_{n,K_n}$ and Σ are as in (2.7) and (2.9), respectively.

Theorem 3.2 shows that the existing asymptotic theory for DML2, where K was fixed as $n \to \infty$, continues to be valid for any K in $\{2, \ldots, n\}$. In particular, we can use DML2 with $K_n = n$, which is exactly the leave-one-out estimator commonly use in semiparametric models as in Robinson (1988), Linton (1995), Rothe and Firpo (2019), among others. In contrast, we cannot use DML1 with $K_n = n$ since the estimator will not be consistent, with some exceptions; see Remark 2.1.

One of the main benefits of using DML2 with $K_n = n$ is that it ensures replicability. DML2 with $K_n = n$ is uniquely determined by the data and therefore eliminates the random-split variability that exists when $K_n < n$, where different random splits yield different DML2 estimates. However, its implementation in practice may appear challenging due to the computational burden of estimating n first-step estimators. Section 5.2 proposes a computationally simple procedure for implementing DML2 with $K_n = n$.

An important caveat is that our results so far do not yet provide guidelines for the choice of K. Our first-order asymptotic theory demonstrates that K does not matter for approximating the finite-sample distribution of DML2 under our assumptions. Therefore, in Section 4, we derive a second-order asymptotic approximation to explain DML2's finite-sample bias and MSE, following a long tradition in econometrics of using second-order asymptotic approximations to compare estimators that are first-order asymptotically equivalent (Rothenberg, 1984; Linton, 1995; Donald and Newey, 2001; Newey and Smith, 2004).

Remark 3.1. A Neyman orthogonality condition of the moment function is necessary to guarantee a first-order equivalent condition between a feasible estimator and its oracle version (as in (2.10)). To see this, consider the following example. Suppose $\psi^a(W, \eta) = 1$ and $\psi^b(W, \eta)$ is a linear function in η. In addition, assume that the nuisance parameter η_0 is an unknown finite-dimensional parameter. Consider the estimator $\hat{\theta}_n = n^{-1}\sum_{i=1}^n \psi^b(W_i, \hat{\eta})$ and its oracle version $\hat{\theta}^*_n = n^{-1}\sum_{i=1}^n \psi^b(W_i, \eta_0)$, where $\hat{\eta}$ is an estimator of η_0 such that

$n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, V_\eta)$ and V_η is an invertible matrix. It can be shown that

$$n^{1/2}(\hat{\theta}_n - \hat{\theta}_n^*) = n^{1/2}(\hat{\eta} - \eta_0)^\intercal E[\partial_\eta m(W_i, \theta_0, \eta_0)] + o_p(1) \ ,$$

which implies $n^{1/2}(\hat{\theta}_n - \hat{\theta}_n^*)$ is $o_p(1)$ if and only if $E[\partial_\eta m(W_i, \theta_0, \eta_0)] = 0$. In other words, the first-order equivalence condition in this example holds if and only if a Neyman orthogonality condition as in part (c) of Assumption 3.1 holds. See also Andrews (1994, Eq. (2.12)). □

Remark 3.2. Moment functions satisfying a Neyman orthogonality condition can be obtained by adding adjustment terms to the original moment functions. The adjustment terms are constructed using first-order influence functions, as developed in Newey (1994), Hahn and Ridder (2013), Ichimura and Newey (2022), and Farrell et al. (2025), among others. Since the analytical construction can be tedious, recent work has focused on automatic construction of orthogonal moments—that is, procedures that take the original moment function as input and automatically return the orthogonalized version needed for DML2 estimation. Examples include Chernozhukov et al. (2022a), Escanciano and Pérez-Izquierdo (2023), and Argañaraz (2025). □

Remark 3.3. Beyond smooth moment conditions, DML methods have been successfully applied to non-smooth econometric models. Chernozhukov et al. (2022a) develop DML estimators for quantile regression coefficients, while Semenova (2023b) propose methods for support functions in set-identified models. Related approaches have been used to study algorithmic fairness (Liu and Molinari, 2025; Liu et al., 2026) and to conduct inference on welfare under optimal treatment rules (Park, 2024). □

3.1 Why and When DML1 is Sensitive to K Increasing

We now provide a high-level explanation for DML1's sensitivity to large K values. The main reason is that the oracle version of DML1 is already sensitive to large K values when the discrepancy measure $\Lambda \neq 0$. Therefore, the discussion that we provided for smooth moment conditions continues to apply for non-smooth econometric models as long as DML1 and its oracle version are asymptotically equivalent.

The oracle version of DML1 is defined as the calculation of DML1 using the true values for η_i's instead of $\hat{\eta}_i$, that is, assuming perfect knowledge of η_0,

$$\hat{\theta}_{n,K}^{*,(1)} = K^{-1} \sum_{k=1}^{K} \tilde{\theta}_k^* \ . \tag{3.2}$$

where

$$\tilde{\theta}_k^* = \left(n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^a(W_i, \eta_i) \right)^{-1} \left(n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^b(W_i, \eta_i) \right) \quad , \quad \forall \ k \in [K].$$

Note that $\tilde{\theta}_k^*$ is similar to the ideal estimator (2.4) but using only observations in the fold $\mathcal{I}_k$.

The next lemma presents the limiting distribution of the oracle DML1 under our new asymptotic framework, in which $K = K_n \to \infty$ as $n \to \infty$.

Lemma 3.1. *Let Assumption 3.1 (a)–(b) hold and let K_n be such that $K_n \leq n$ and $K_n/\sqrt{n} \to c \in [0, \infty)$ as $n \to \infty$. Then,*

$$\sqrt{n} \left(\hat{\theta}_{n,K_n}^{*,(1)} - \theta_0 \right) \xrightarrow{d} N(c\Lambda, \Sigma) \ ,$$

where Σ and Λ are defined in (2.9) and (3.1), respectively.

Lemma 3.1 continue to hold for non-smooth econometric models for two reasons. First, we use mild regularity conditions to derive the limiting distribution of the oracle DML1. Second, we don't rely on the smoothness of the moment function m with respect to η.

Lemma 3.1 shows that the oracle DML1 has the same limiting distribution that we derive for DML1 in Theorem 3.1. This last result occurs because the proof of Theorem 3.1 uses that the DML1 and its oracle version are asymptotically equivalent,

$$\sqrt{n} \left(\hat{\theta}_{n,K_n}^{(1)} - \hat{\theta}_{n,K_n}^{*,(1)} \right) \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty \ . \tag{3.3}$$

Part (d) of Assumption 3.1 is key in our proof of Theorem 3.1 to guarantee that (3.3) holds.

The asymptotic equivalence in (3.3) and Assumption 3.1 (a)-(b) are sufficient to conclude that DML1 is sensitive for large K values when $\Lambda \neq 0$. We show in part (a) of Lemma A.1 that Assumptions 3.1 and 3.2 are sufficient to verify the high-level condition (3.3).

In a similar way, we can define the oracle version of DML2:

$$\hat{\theta}_{n,K}^{*,(2)} = \left(\frac{1}{n} \sum_{i=1}^{n} \psi^a(W_i, \eta_i) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \psi^b(W_i, \eta_i) \right) \ . \tag{3.4}$$

Note that the oracle version of the DML2 is the same as the ideal estimator defined in (2.4). Therefore, the oracle version of DML2 does not depend on the choice of K.

The asymptotic equivalence in (3.5) between DML2 and its oracle version is sufficient to

conclude the robustness of DML2 to the choice of K,

$$\sqrt{n}\left(\hat{\theta}^{(2)}_{n,K_n} - \hat{\theta}^{*,(2)}_{n,K_n}\right) \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty \,. \tag{3.5}$$

We show in part (b) of Lemma A.1 that Assumptions 3.1 (a)–(c) and 3.2 are sufficient to verify the high-level condition (3.5), provided that $K_n = O(\sqrt{n})$. To guarantee that (3.5) holds for $K_n = O(n)$ we additionally use Assumption 3.3 (Lemma A.2). Importantly, we do not require part (d) of Assumption 3.1 for the analysis of DML2.

Finally, notice that the oracle version of DML1 depends on random splitting, while the oracle version of DML2 does not. Therefore, even if we use the oracle DML1, it lacks replicability due to sample splitting, whereas the oracle DML2 does not. Furthermore, we cannot use oracle DML1 with $K = n$ since this estimator is inconsistent, except in special cases where oracle DML1 coincides with oracle DML2; see Remark 2.1.

4 Second-Order Asymptotic Approximation for DML2 when K increases

This section derives a second-order asymptotic approximation to the scalar DML2 estimator when $K = K_n$ can depend on the sample size n (Theorem 4.1). We use this approximation to explain observed patterns in DML2's finite-sample bias and MSE (Remark 2.3), characterize the optimal choice of K, and quantify the relative efficiency loss from any suboptimal choice of K. Under the conditions we provide, the magnitude of the second-order asymptotic bias and the second-order asymptotic MSE of DML2 decrease in K, implying that $K_n = n$ is an optimal choice. In other words, the leave-one-out estimator is optimal among DML2 estimators; its implementation is discussed in Section 5.2.

Let $n_0^{-\varphi}$ be the L_2-convergence rate of the estimator $\hat{\eta}_k$, where $n_0 = ((K-1)/K)n$ is the number of observations in the sample $\{W_i : i \notin \mathcal{I}_k\}$ used by $\hat{\eta}_k$ to estimate η_0. Let M_1, M_2, c_δ, and c_b be positive constants, and let τ_n be a sequence of positive numbers converging to zero. To derive the second-order asymptotic approximation for DML2, the next assumption restricts the class of nuisance estimators relative to that considered in Section 3, in the sense that it can be verified that Assumption 4.1 implies Assumptions 3.2 and 3.3.

Assumption 4.1. *There exist $\delta_{n_0} : \mathcal{W} \times \mathcal{X} \to \mathbf{R}^p$ and $b_{n_0} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}^p$, such that*

(a) For any $k \in [K_n]$, $E[||\hat{\eta}_k(X_i) - \eta_0(X_i) - \Delta_{k,i}||^2]^{1/2} \leq n_0^{-2\varphi} M_1$, where

$$\Delta_{k,i} = n_0^{-1/2} \sum_{j \notin \mathcal{I}_k} n_0^{-\varphi} \delta_{n_0}(W_j, X_i) + n_0^{-1} \sum_{j \notin \mathcal{I}_k} n_0^{-\varphi} b_{n_0}(X_j, X_i) \,. \tag{4.1}$$

(b) $\varphi \in (1/4, 1/2)$.

(c) For any $i \neq j$, $E\left[||\delta_{n_0}(W_j, X_i)||^2\right] \in (c_\delta, M_1)$, $E\left[||\delta_{n_0}(W_j, X_i)||^4\right] < n_0^{1-2\varphi} M_2$ *for* $s = 1, 2$, $E\left[E[||\delta_{n_0}(W_j, X_i)||^2 \mid X_i]^2\right] \leq M_2$, *and* $E[\delta_{n_0}(W_j, X_i) \mid X_j] = 0$.

(d) For any $i \neq j$, $E\left[||E\left[b_{n_0}(X_j, X_i) \mid X_i\right]||^4\right] \in (c_b, M_2)$, $E\left[||n_0^{-\varphi} b_{n_0}(X_j, X_i)||^{2s}\right] < n_0^{(2s-1)(1-2\varphi)} \tau_{n_0}$ *for* $s = 1, 2$, *and* $E\left[E\left[||b_{n_0}(X_j, X_i)||^2 \mid X_i\right]^2\right] \leq n_0^2 \tau_{n_0}$.

Assumption 4.1 holds for kernel regression estimators under mild regularity conditions, with trimming to ensure bounded inverse density weights. It captures settings in which the nuisance estimator achieves the usual bias–variance trade-off. For instance, it holds for the Nadaraya–Watson estimator with an MSE-optimal bandwidth.

Part (a) of Assumption 4.1 is a high-level condition that presents a stochastic expansion for the nuisance function estimator $\hat{\eta}_k$, with variance and bias contributions given by δ_{n_0} and b_{n_0}, respectively. The accuracy of the stochastic expansion is measured in L_2-norm to address the technical challenges that arise when $K_n \to \infty$ as $n \to \infty$; it can be relaxed when K_n is fixed. Part (b) of Assumption 3.1 is a standard requirement in the semiparametric literature (Andrews, 1994); it guarantees that nonlinear effects of first-step estimator error are negligible for the estimator of θ_0 whenever part (c) of Assumption 3.1 also holds. Parts (c) and (d) impose regularity conditions on δ_{n_0} and b_{n_0} that imply $n_0^{-2\varphi}$ is the convergence rate of both the squared bias and variance of $\hat{\eta}_k$.

Assumption 4.1 provides additional structure on the estimators $\hat{\eta}_k$ that we can use to derive a second-order asymptotic approximation for the scalar DML2 estimator. Nevertheless, conducting an appropriate analysis of the leading terms of the second-order bias and MSE of DML2 requires additional conditions on the functions δ_{n_0} and b_{n_0} and the higher-order partial derivatives of the moment function m. We formalize those conditions in the next assumption. To simplify notaton, let $\tilde{b}_{n_0}(X_i) = E\left[b_{n_0}(X_j, X_i) \mid X_i\right]$ for $j \neq i$, let $\partial_\eta^2 m_i = \partial_\eta^2 m(W_i, \theta_0, \eta_i)$, and recall that $\eta_i = \eta_0(X_i)$ and $G = E[\psi^a(W_i, \eta_i)]$.

Assumption 4.2. *(a)* m *is three-times continuously differentiable on* $\eta \in \mathcal{E} \subseteq \boldsymbol{R}^p$ *and* $\sup_{\eta \in \mathcal{E}} ||\partial_\eta^3 m(W_i, \theta_0, \eta)||_\infty \leq c_2$, *(b) the next limits exist and are finite,*

$$\mathcal{V} = \lim_{n_0 \to \infty} E\left[E\left[\delta_{n_0}(W_j, X_i)^\intercal \left(G^{-1} \partial_\eta^2 m_i\right) \delta_{n_0}(W_\ell, X_i) \mid W_j, W_\ell\right]^2\right], \tag{4.2}$$

$$\mathcal{B} = \lim_{n_0 \to \infty} \tfrac{1}{2} E\left[\left(\delta_{n_0}(W_j, X_i) + \tilde{b}_{n_0}(X_i)\right)^\intercal \left(G^{-1} \partial_\eta^2 m_i\right) \left(\delta_{n_0}(W_j, X_i) + \tilde{b}_{n_0}(X_i)\right)\right] \tag{4.3}$$

$$\mathcal{C} = \lim_{n_0 \to \infty} E\left[G^{-1} m(W_j, \theta_0, \eta_j) \delta_{n_0}(W_j, X_i)^\intercal \left(G^{-1} \partial_\eta^2 m_i\right) \tilde{b}_{n_0}(X_i)\right], \tag{4.4}$$

where $j \neq i$, *and (c)* $\mathcal{V} > 0$, $\mathcal{B} \neq 0$, *and* $\mathcal{C} > 0$.

Part (a) of Assumption 4.2 is satisfied in several examples, including ATE (Example C.1), ATT-DID (Example C.2), and LATE (Example C.3). In all these examples, the moment function is a quadratic polynomial in η. Part (b) requires that the limits in (4.2)–(4.4) exist; this is a mild regularity condition since Assumptions 3.1 and 4.1 already ensure these sequences are bounded. Part (c) assumes that the quantities $\mathcal{V}$, $\mathcal{B}$ and $\mathcal{C}$ are non-zero to ensure the second-order approximation is non-degenerate. A necessary condition for part (c) is that the moment function m is nonlinear in η, i.e., its matrix of second-order partial derivatives with respect to η is nonzero.

Let $\mathcal{T}_n^* = n^{-1/2}\sum_{n=1}^{n} G^{-1}m_i$ and $\mathcal{T}_{n,K}^{nl} = \frac{1}{2}n^{-1/2}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\Delta_{k,i}^{\intercal}\left(G^{-1}\partial_\eta^2 m_i\right)\Delta_{k,i}$, where we use $m_i = m(W_i,\theta_0,\eta_i)$ to simplify notation. Recall that $\Delta_{k,i}$ is defined in (4.1). We refer $\mathcal{T}_n^*$ as the first-order asymptotic approximation of DML2 since $\mathcal{T}_n^* \xrightarrow{d} N(0,\Sigma)$, which is the limiting distribution of DML2. Let $\mathcal{T}_{n,K_n} \equiv \mathcal{T}_n^* + \mathcal{T}_{n,K_n}^{nl}$. The next theorem shows that $\mathcal{T}_n^*$ and $\mathcal{T}_{n,K_n}$ are, respectively, the first- and second-order asymptotic approximations to the scalar DML2 estimator.

Theorem 4.1. *Let Assumptions 3.1, 4.1, and 4.2 hold and let K_n be such that $K_n \leq n$. Then,*

$$n^{1/2}(\hat{\theta}_{n,K_n}^{(2)} - \theta_0) - \mathcal{T}_{n,K_n} = o_p(n^{1/2-2\varphi}) \ . \tag{4.5}$$

Furthermore, $\lim_{n\to\infty} Var[n^{2\varphi-1/2}\mathcal{T}_{n,K_n}^{nl}] > 0$ and $\lim_{n\to\infty} E\left[\left(n^{2\varphi-1/2}\mathcal{T}_{n,K_n}^{nl}\right)^2\right] < \infty$. In particular,

$$n^{1/2}(\hat{\theta}_{n,K_n}^{(2)} - \theta_0) - \mathcal{T}_n^* = O_p(n^{1/2-2\varphi}) \ .$$

Theorem 4.1 demonstrates that $\mathcal{T}_{n,K_n}$ provides a better asymptotic approximation than $\mathcal{T}_n^*$. We obtain this improvement by including $\mathcal{T}_{n,K_n}^{nl}$ to account for the nonlinear effects of nuisance estimation error in the estimator of θ_0. More concretely, the theorem guarantees that $\mathcal{T}_{n,K_n}^{nl}$ has stochastic order $O_p(n^{1/2-2\varphi})$ and is the leading term in the scaled difference between the feasible estimator $\hat{\theta}_{n,K_n}^{(2)}$ and the oracle estimator $\hat{\theta}_{n,K_n}^{*,(2)}$ defined in (3.4):

$$n^{1/2}\left(\hat{\theta}_{n,K_n}^{(2)} - \hat{\theta}_{n,K_n}^{*,(2)}\right) = \mathcal{T}_{n,K_n}^{nl} + o_p(n^{1/2-2\varphi}).$$

We next use the asymptotic approximation $\mathcal{T}_{n,K_n}$ to explain how the choice of K_n affects the finite-sample bias and MSE of DML2. Recall that all DML2 estimators share the same limiting distribution regardless of K_n (Theorem 3.2), making second-order analysis necessary to understand the simulation patterns of the bias and MSE of DML2 that first-order asymptotic theory cannot capture. This use of second-order approximations to compare first-order equivalent estimators has a long history in econometrics (Rothenberg, 1984; Linton, 1995; Newey and Smith, 2004; Graham et al., 2012).

Remark 4.1. When K_n is fixed as $n \to \infty$, a similar stochastic expansion can be derived for DML1,

$$n^{1/2}(\hat{\theta}^{(1)}_{n,K_n} - \theta_0) = \mathcal{T}^*_n + \mathcal{T}^{nl}_{n,K_n} + o_p(n^{1/2-2\varphi}) \ .$$

This expression and (4.5) show that the asymptotic approximations are identical when K_n is fixed. Consequently, DML1 and DML2 have identical second-order asymptotic bias and MSE in the K-fixed asymptotic regime, making it impossible to distinguish them using second-order asymptotic analysis. Thus, distinguishing DML1 from DML2 requires an asymptotic framework where $K_n \to \infty$ as $n \to \infty$, as we develop in Section 3. □

4.1 Second-order Asymptotic bias and MSE for DML2

We define the second-order asymptotic bias and MSE of DML2 as the mean and second moment of $\mathcal{T}_{n,K_n}$, respectively. These definitions follow a long tradition in econometrics of using second-order approximations to compare estimators with identical first-order asymptotic properties (Rothenberg, 1984; Linton, 1995; Newey and Smith, 2004). Under suitable regularity conditions, the distribution of $\mathcal{T}_{n,K_n}$ approximates the distribution of $n^{1/2}(\hat{\theta}^{(2)}_{n,K_n} - \theta_0)$ up to an error of order $o(n^{1/2-2\varphi})$; therefore, the moments of $\mathcal{T}_{n,K_n}$ provide valid approximations to the bias and variance of $\hat{\theta}^{(2)}_{n,K_n}$. Recall that $\varphi \in (1/4, 1/2)$ by Assumption 4.1.

Theorem 4.2. *Let Assumptions 3.1, 4.1, and 4.2 hold and let K_n be such that $K_n \leq n$. Then,*

$$E[\mathcal{T}_{n,K_n}] = \mathcal{B}\left(1 + \frac{1}{K_n - 1}\right)^{2\varphi} n^{1/2-2\varphi} + o(n^{1/2-2\varphi}) \ , \tag{4.6}$$

and

$$E[\mathcal{T}^2_{n,K_n}] = \Sigma + \mathcal{C}\left(1 + \frac{1}{K_n - 1}\right)^{2\varphi-1/2} n^{1/2-2\varphi} + o(n^{1/2-2\varphi}) \ , \tag{4.7}$$

where $\mathcal{B}$ and $\mathcal{C}$ are defined in (4.2) and (4.4), respectively.

This theorem presents the second-order asymptotic bias and MSE of DML2. Henceforth, by second-order asymptotic bias and MSE we refer to the leading-order terms in (4.6) and (4.7), omitting the $o(n^{1/2-2\varphi})$ terms, which are negligible for our analysis. This simplification focuses our analysis on the dominant terms that vary with K_n.

Theorem 4.2 provides an asymptotic result that explains the observed patterns in DML2's finite-sample bias and MSE (Remark 2.3). Since $\mathcal{B} \neq 0$ and $\mathcal{C} > 0$, we see that the magnitude of the second-order asymptotic bias and the second-order asymptotic MSE decrease in K_n, consistent with the simulations findings reported in Ahrens et al. (2024a,b) and Chernozhukov et al. (2018), and our simulation results in Section 6. Importantly, the simulation results in Section 6 show that for $K \geq 10$, DML2's finite-sample bias and MSE

appear approximately constant. This plateau is consistent with the fact that the terms $(1-1/(K_n-1))^{2\varphi}$ and $(1-1/(K_n-1))^{2\varphi-1/2}$ in (4.6)–(4.7) change little when $K_n \geq 10$ for typical values $\varphi \in (1/4, 1/2)$.

We now use Theorem 4.2 to characterize the optimal choice of $K = K_n$, which can depend on the sample size n. We consider the minimization of the second-order asymptotic MSE of DML2 as our optimality criterion, following the literature on higher-order asymptotics (Donald and Newey, 2001; Linton, 1995; Newey and Smith, 2004). Let $MSE[\hat{\theta}^{(2)}_{n,K_n}]$ be the second-order MSE of DML2 with $K = K_n$. Using this notation and since $\mathcal{C} > 0$, we conclude

$$MSE[\hat{\theta}^{(2)}_{n,n}] \geq MSE[\hat{\theta}^{(2)}_{n,K_n}] \tag{4.8}$$

for any sequence K_n such that $K_n \leq n$.

From (4.8), we conclude that $K = n$ is an optimal choice for DML2 in terms of second-order asymptotic MSE. When K_n is constant, the inequality (4.8) is strict, implying that $K = n$ strictly dominates any fixed choice of K. In contrast, when $K_n \to \infty$ as $n \to \infty$, the difference between $MSE[\hat{\theta}^{(2)}_{n,n}]$ and $MSE[\hat{\theta}^{(2)}_{n,K_n}]$ is of order $o(n^{1/2-2\varphi})$, which we omit in our analysis. Thus, any choice K_n with $K_n \to \infty$ as $n \to \infty$ is asymptotically equivalent to $K = n$ under the second-order asymptotic MSE criterion.

The previous result demonstrate that the leave-one-out estimator, which is DML2 with $K = n$, is optimal among DML2's in terms of second-order asymptotic MSE. A similar analysis can be done using the magnitude of the second-order asymptotic bias of DML2 as our optimality criterion, with analog results, in the sense that the choice $K = n$ is also an optimal choice as long as $\mathcal{B} \neq 0$. Therefore, the leave-one-out estimator is also optimal among DML2's in terms of second-order asymptotic bias.

Remark 4.2. Under our conditions (including $\mathcal{B} \neq 0$ and $\mathcal{C} > 0$), we prove that $K_n = n$ is optimal in terms of second-order asymptotic bias and MSE criteria. If instead $\mathcal{C} < 0$, then $K_n = 2$ becomes optimal under the second-order asymptotic MSE criterion. In either case, commonly recommended choices as $K = 5$ or $K = 10$ are suboptimal. □

Remark 4.3. When $K_n = K$ is fixed as $n \to \infty$, explicit expressions for the second-order asymptotic MSE of the oracle estimators $\hat{\theta}^{*,(1)}_{n,K}$ and $\hat{\theta}^{*,(2)}_{n,K}$ defined in (3.2) and (3.4), respectively, can be derived using standard arguments (e.g., Newey and Smith (2004)):

$$MSE[\hat{\theta}^{*,(1)}_{n,K}] = \Sigma + \left(K^2\Lambda^2 + K\Lambda_1\right)/n + o(n^{-1})$$
$$MSE[\hat{\theta}^{*,(2)}_{n,K}] = \Sigma + \left(\Lambda^2 + \Lambda_1\right)/n + o(n^{-1})$$

where Σ and Λ are defined in (2.9) and (3.1), respectively, and

$$\Lambda_1 = 5\Lambda^2 + \sigma^2 \left\{ 3\frac{E\left[\psi^a(W,\eta_0(X))^2\right]}{E\left[\psi^a(W,\eta_0(X))\right]^2} - 1 \right\} - 2\frac{E\left[m(W,\theta_0,\eta_0(X))^2\psi^a(W,\eta_0(X))\right]}{E\left[\psi^a(W,\eta_0(X))\right]^3} \ .$$

Two key differences from Theorem 4.2 merit discussion. First, the oracle estimators have remainder terms of order $o(n^{-1})$ and second-order terms of order n^{-1}. Since $\varphi \in (1/4, 1/2)$ implies $n^{-1} = o(n^{1/2-2\varphi})$, these oracle second-order terms are negligible relative to the feasible estimator's second-order MSE term in (4.7). Second, the second-order term for DML1, $\left(K^2\Lambda^2 + K\Lambda_1\right)/n$, increases in K, implying that for large K, the oracle DML1 estimator $\hat{\theta}^{*,(1)}_{n,K}$ has worse second-order accuracy than the oracle DML2 estimator $\hat{\theta}^{*,(2)}_{n,K}$. □

Remark 4.4. Theorem 4.2 illustrates that a first-step estimator optimal for nuisance estimation may be suboptimal for estimating θ_0 in terms of second-order asymptotic MSE. The theorem shows that the second-order asymptotic MSE of DML2 is dominated by the variance component since the squared second-order bias is of order $O(n^{1-4\varphi})$, which is a negligible relative to the variance. Therefore, a different class of first-step estimators that induces larger second-order bias but lower second-order variance could improve the convergence rate of DML2's second-order MSE. In work in progress, we show that when the first-step estimator uses Nadaraya-Watson regression, the bandwidth $h_n \propto n^{-2/7}$ optimizes the second-order MSE of $\hat{\theta}^{(2)}_{n,K_n}$, yielding a convergence rate of $n^{-3/7}$, which is faster than the rate $n^{-3/10}$ obtained in (4.7) using the MSE-optimal bandwidth $h_n \propto n^{-1/5}$. □

4.2 Relative efficiency loss from suboptimal choice of K

In the remainder of this section, we quantify the relative efficiency loss from any suboptimal choice of K using Theorem 4.2. The main motivation is to evaluate the performance of commonly recommended choices such as $K = 5$ or $K = 10$ relative to the optimal choice. We first use the second-order asymptotic MSE as our performance metric, then present results using the second-order asymptotic bias.

We define the relative efficiency loss of the choice K in terms of the second-order asymptotic MSE as

$$\mathcal{RL}_{\text{MSE}}(K) \equiv \frac{MSE[\hat{\theta}^{(2)}_{n,K}]}{MSE[\hat{\theta}^{(2)}_{n,n}]} - 1 \ .$$

This measures the percentage loss in second-order asymptotic MSE from choosing K instead of the optimal $K = n$. By construction, $\mathcal{RL}_{\text{MSE}}(n) = 0$ and $\mathcal{RL}_{\text{MSE}}(K) \geq 0$ for all $K \leq n$.

The next corollary provides an upper bound for $\mathcal{RL}_{\text{MSE}}(K)$ depending only on K, n, and φ. This bound is sufficiently tight for practical guidance and avoids the need to estimate the ratio $\mathcal{C}/\Sigma$ required in the exact expression.

Corollary 4.1. *Under the conditions of Theorem 4.2, we have*

$$\mathcal{RL}_{MSE}(K) \leq \frac{\left(1+\frac{1}{K-1}\right)^{2\varphi-1/2}}{\left(1+\frac{1}{n-1}\right)^{2\varphi-1/2}} - 1 \ .$$

In particular, if $\varphi \in (1/4, 1/2)$*, we have* $\mathcal{RL}_{MSE}(5) \leq 11.8\%$ *and* $\mathcal{RL}_{MSE}(10) \leq 5.4\%$ *for* $n \geq 1000$*. If we know* $\varphi = 2/5$*, we have* $\mathcal{RL}_{MSE}(5) \leq 6.9\%$ *and* $\mathcal{RL}_{MSE}(10) \leq 3.2\%$*.*

This corollary shows that the relative efficiency loss from commonly recommended choices such as $K = 5$ or $K = 10$ is small. Moreover, these relative losses decrease as the first-step estimator becomes less accurate (i.e., as φ decreases), indicating that optimal choice of K is less critical when nuisance estimation is slower.

We now define the relative efficiency loss of the choice K in terms of the second-order asymptotic bias as

$$\mathcal{RL}_{\text{bias}}(K) \equiv \left(\frac{1+\frac{1}{K-1}}{1+\frac{1}{n-1}}\right)^{2\varphi} - 1 \ . \tag{4.9}$$

This measures the percentage loss in second-order asymptotic bias from choosing K instead of the optimal one. By construction, $\mathcal{RL}_{\text{bias}}(n) = 0$ and $\mathcal{RL}_{\text{bias}}(K) \geq 0$ for all $K \leq n$. Recall that we are referring to the second-order asymptotic bias to the leading-order term in (4.6), since the terms of order $o(n^{1/2-2\varphi})$ are negligible for our analysis.

From (4.9), we conclude that $\mathcal{RL}_{\text{bias}}(5) \in (11.7\%, 24.9\%)$ and $\mathcal{RL}_{\text{bias}}(10) \in (5.4\%, 11\%)$ when $\varphi \in (1/4, 1/2)$ and $n \geq 1000$; therefore, the relative efficiency loss from commonly recommended choices such as $K = 5$ or $K = 10$ may be significative in terms of second-order asymptotic bias. Furthermore, these relative losses increases as the first-step estimator becomes more accurate (i.e., as φ increases), showing that optimal choice of K is critical when nuisance estimation is faster.

5 Recommendations for Practitioners

We now provide three recommendations for the implementation of DML. In contrast to existing guidance, which relies on simulation evidence that necessarily cannot cover the full range of econometric models and first-step estimators, our recommendations are based on the theoretical results presented in Sections 3 and 4.

Before presenting our recommendations, we recall that DML offers practitioners multiple implementation choices. These include the choice between DML1 and DML2, as well as the number of folds K used to split the data for first-step estimation. Our recommendations focus on these two key decisions: DML1 versus DML2, and the choice of K.

5.1 Prefer DML2 over DML1

Our first recommendation for practitioners is to use DML2 over DML1. While this is consistent with existing practice, we provide theoretical justification that was previously lacking.

We offer two supporting reasons. First, DML2 asymptotically dominates DML1 in terms of both bias and MSE (Theorem 3.1). Moreover, standard inference based on DML1 is invalid when $\Lambda \neq 0$ (Corollary 3.1), where Λ is the discrepancy measure defined in (3.1)—a quantity that can be computed without data. In contrast, DML2-based inference remains valid (Corollary 3.2). Second, DML2 is robust to the choice of K for a large class of first-step estimators (Theorem 3.2). In other words, estimation and inference using DML2 are reliable for any choice of $K \in \{2, \ldots, n\}$.

Importantly, DML1 can still be used when $\Lambda = 0$ and $K \propto \sqrt{n}$, since under these conditions DML1 achieves the same first-order asymptotic properties as DML2. However, this requires first calculating Λ and verifying whether it equals zero. In practice, it is simpler to use DML2 directly. Moreover, implementations of DML2 are available for many econometric models in `Stata` (*ddml*; Ahrens et al. (2024a)), `Python` (*DoubleML*; Bach et al. (2022)), and `R` (*DoubleML*; Bach et al. (2024)).

5.2 Use $K = n$ for DML2

Our second recommendation for practitioners is to use DML2 with $K = n$. There are two reasons supporting this choice. First, $K = n$ is optimal for DML2 in terms of second-order asymptotic bias and MSE, as we show in Section 4.1. Second, $K = n$ ensures replicability by eliminating random-split variability. DML2 with $K = n$ is uniquely determined by the data. In contrast, for any $K < n$, different random splits yield different DML2 estimates, so researchers analyzing the same data with the same K could obtain different conclusions.

We now propose a computationally simple procedure to implement DML2 with $K = n$. We start by recognizing that the standard practice in DML is to repeat the tuning of the first-step estimator K times. Here, by tuning, we refer to the procedure in which hyperparameters are selected—such as the bandwidth in kernel regression or penalization parameters in LASSO—before estimation. Tuning is often implemented via cross-validation (Hastie et al., 2009), which is computationally demanding and uses the same data that will later be used for estimation.

Instead of following the standard practice, we propose to tune the first-step estimator only once using all the data. This will result in a unique set of hyperparameters that we use to estimate all the first-step estimators. This procedure relies on the assumption that a single observation has negligible influence on the selected hyperparameters—an assump-

tion that seems reasonable in practice. Under this assumption, the tuning procedures for different first-step estimators—each using $n-1$ observations since $K=n$—will yield nearly identical hyperparameters, as they differ by only one observation. Moreover, these hyperparameters will be close to those obtained using the full dataset. Therefore, our proposed procedure provides a computationally simpler alternative to the standard practice, though formal verification of this assumption remains for future work.

5.3 If K must be small, use $K=10$ over $K=5$

When practitioners must choose a small value of K for DML2, they should use $K=10$ over $K=5$. The reason is that DML2 with $K=10$ achieves better second-order asymptotic accuracy than $K=5$, as we show in Section 4.1. Moreover, the relative efficiency loss from choosing $K=5$ versus the optimal $K=n$ can be as high as 11.8% in terms of second-order MSE, while choosing $K=10$ reduces this to at most 5.4% (Corollary 4.1). Therefore, $K=10$ guarantees substantially lower efficiency losses than $K=5$.

Finally, Section 4.2 presents simple formulas to calculate the relative efficiency loss from suboptimal choices of K. See (4.9) and Corollary 4.1 for the relative efficiency loss in terms of second-order asymptotic bias and MSE, respectively.

6 Simulations

This section examines how well the asymptotic approximations from Sections 3 and 4 capture finite-sample behavior for two econometric models: (i) ATT-DID (Sant'Anna and Zhao, 2020) and (ii) LATE (Hong and Nekipelov, 2010). We calculate the bias, MSE, and coverage probability of confidence intervals associated with DML1 and DML2 for several values of K. We use the confidence intervals defined in Chernozhukov et al. (2018, Theorem 3.2). In what follows, we first present the designs and then the simulation results.

6.1 Design: LATE and ATT-DID

6.1.1 ATT-DID

This section is based on Example C.2. We built on the simulation design presented in Sant'Anna and Zhao (2020). The observed outcome in the pre-treatment period and the potential outcomes in the post-period treatment are defined by

$$Y_{0,i} = f_{reg}(X_i) + v(X_i, A_i) + \varepsilon_{0,i}$$

$$Y_{1,i}(a) = 2f_{reg}(X_i) + v(X_i, A_i) + \varepsilon_{1,i}(a) \ , \quad a = 0, 1$$

where $f_{reg}(X) = 210 + 6.85X_1 + 3.425(X_2 + X_3 + X_4)$ and $v(X_i, A_i) = A_i f_{reg}(X) + \varepsilon_{v,i}$, and $(\varepsilon_{0,i}, \varepsilon_{1,i}(0), \varepsilon_{1,i}(1), \varepsilon_{v,i})$ is distributed as $N(0, \mathbb{I}_4)$, $\mathbb{I}_4$ is the 4×4 identity matrix. The treatment assignment is defined by $A_i \sim \text{Bernoulli}(p(X_i))$, where

$$p(X_i) = \frac{\exp(f_{ps}(X_i))}{1 + \exp(f_{ps}(X_i))}$$
$$f_{ps}(X) = 0.25(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4) \ .$$

Finally, the vector of covariates is $X_i = (X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}) \in [0, 1]^4$ and all its coordinates are independent uniform random variables (e.g., $X_{1,i} \sim \text{Uniform}[0, 1]$).

6.1.2 LATE

This section is based on Example C.3. We built on the simulation design presented in Hong and Nekipelov (2010). The potential treatment decisions are defined as

$$D_i(1) = I\{X_i + 0.5 \geq V_i\} \ ,$$
$$D_i(0) = I\{X_i - 0.5 \geq V_i\} \ ,$$

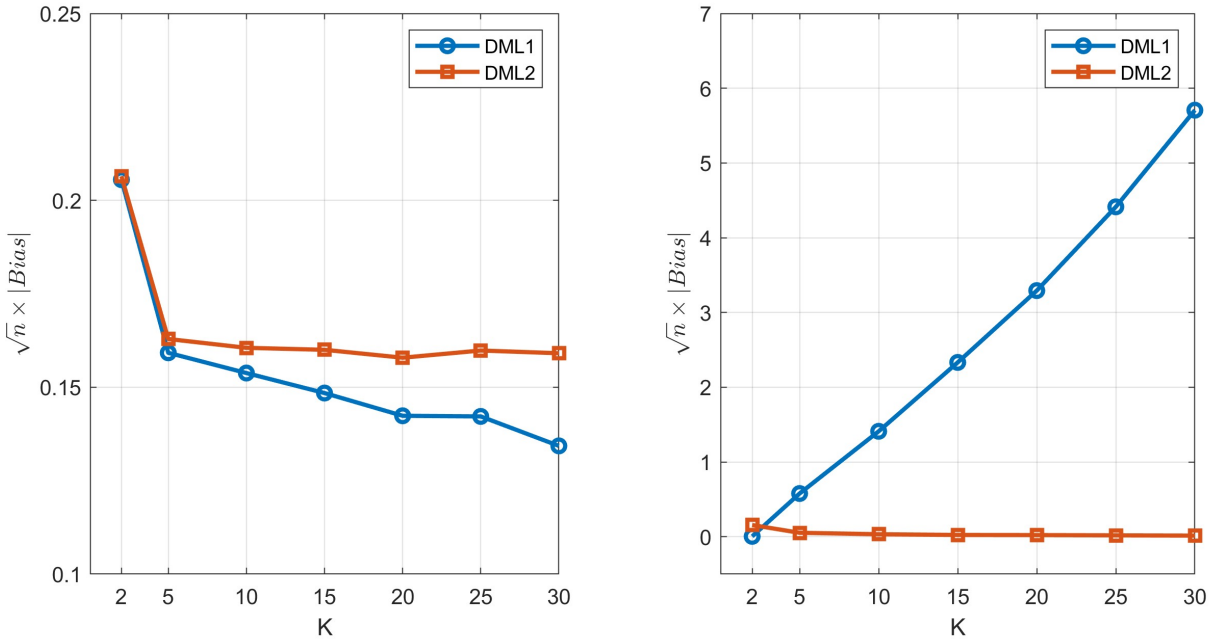
where $X_i \sim \text{Uniform}[0, 1]$ and $V_i \sim N(0, 1)$ are independent random variables. The potential outcomes are defined by

$$Y_i(1) = \xi_{1,i} + \xi_{3,i} I\{D_i(1) = 1, D_i(0) = 1\} + \xi_{4,i} I\{D_i(1) = 0, D_i(0) = 0\} \ ,$$
$$Y_i(0) = \xi_{2,i} + \xi_{3,i} I\{D_i(1) = 1, D_i(0) = 1\} + \xi_{4,i} I\{D_i(1) = 0, D_i(0) = 0\} \ ,$$

where $\xi_{1,i} \sim \text{Poisson}(\exp(X_i/2))$, $\xi_{2,i} \sim \text{Poisson}(\exp(X_i/2))$, $\xi_{3,i} \sim \text{Poisson}(2)$, and $\xi_{4,i} \sim \text{Poisson}(1)$, and all these random variables are independent conditional on X_i. The treatment assignment is defined by $Z_i \sim \text{Bernoulli}(\Phi(X_i - 0.5))$. As in Example C.3, the observed treatment decision and the observed outcome are defined by $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$ and by $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$, respectively.

6.2 Results: LATE is sensitive to K increasing, while ATT-DID is not

This section provides simulation evidence showing that DML2 strictly dominates DML1 in the case of LATE, but performs similarly for the case of ATT-DID. This is consistent with



(a) ATT-DID ($\Lambda = 0$)

(b) LATE ($\Lambda \neq 0$)

Figure 1: Bias of DML1 and DML2 for ATT-DID and LATE. Sample size $n = 3,000$ and 5,000 simulations.

our results in Section 3 since LATE has $\Lambda \neq 0$, while ATT-DID has $\Lambda = 0$. Below we provide additional details on the construction of the DML estimators.

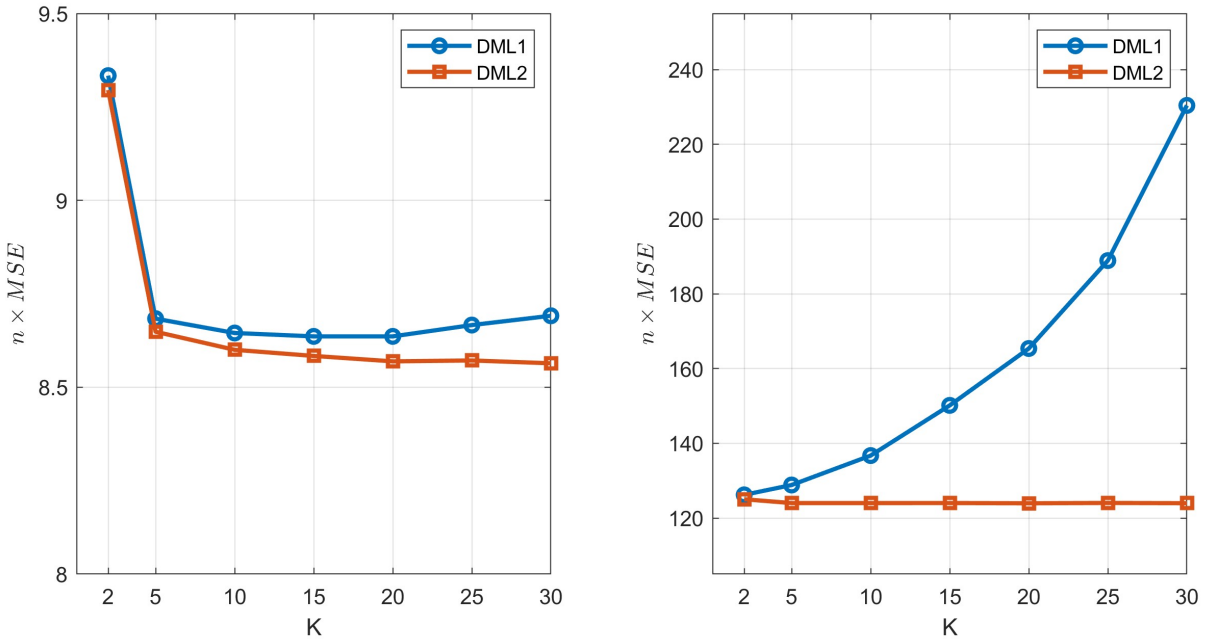
The estimators for the ATT-DID and LATE are defined as in (2.6) and (2.7) using ψ^a and ψ^b presented in Examples C.2 and C.3, respectively. They are calculate for different values of $K \in \{2, 5, 10, 15, 20, 25, 30\}$. The nuisance function η_0 for the ATT-DID and LATE are presented in Examples C.2 and C.3, respectively.

We estimate each component of the nuisance function η_0 using Nadaraya-Watson estimators and the cross-fitting procedure described in Section 2.1, where each first-step estimator uses sample size $n_0 = ((K-1)/K)n$. For the ATT-DID, we use a 6th-order Gaussian kernel and common bandwidth $h_j = cn_0^{-1/16}$ for all coordinates.[3] For the LATE, we use a 2nd-order Gaussian kernel and common bandwidth $h_j = cn_0^{-1/5}$.

6.2.1 Bias

Figure 1 presents the bias of DML1 and DML2 for several values of K and two econometric models: ATT-DID in panel (a) and LATE in panel (b). Panel (a) shows that DML1 and

[3]We also considered a 2nd order Gaussian Kernel in the simulations. The results are presented in Figures D.1 and D.2 in Appendix D, and they are similar to the ones presented using a 6th order Gaussian kernel.



(a) ATT-DID ($\Lambda = 0$)

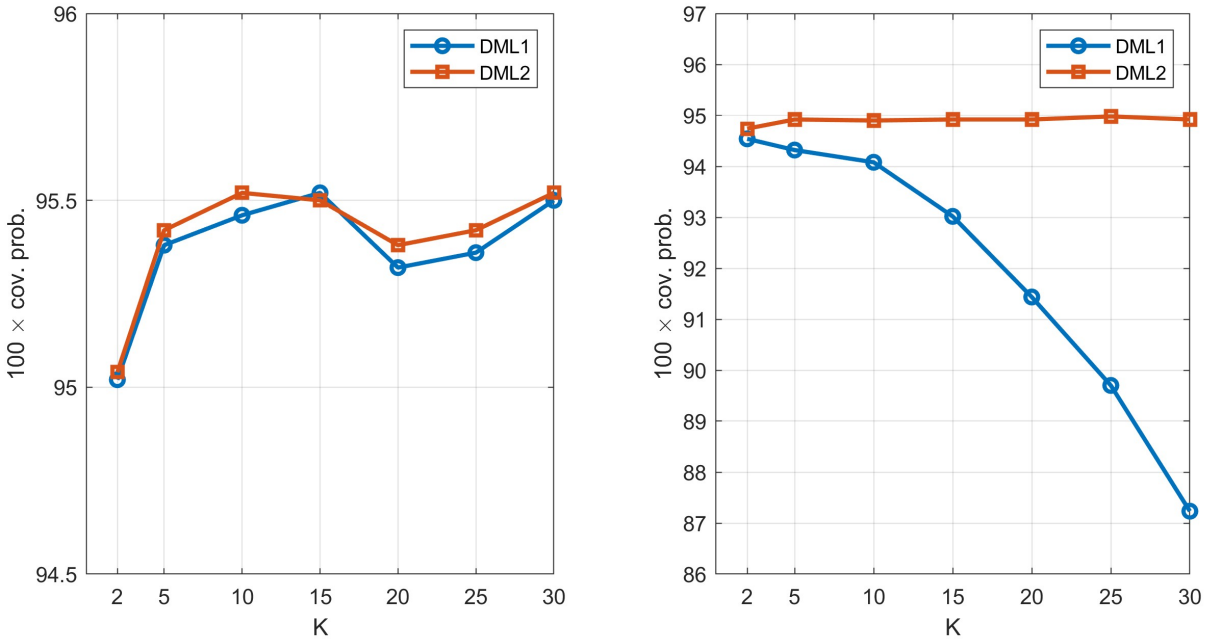
(b) LATE ($\Lambda \neq 0$)

Figure 2: MSE of DML1 and DML2 for ATT-DID and LATE. Sample size $n = 3,000$; 5,000 simulations.

DML2 perform similarly in terms of bias, while panel (b) shows that the bias of DML1 grows almost linearly in K. Theorem 3.1 explains this finite-sample behavior since $\Lambda = 0$ for panel (a) and $\Lambda \neq 0$ for panel (b). Importantly, in both panels, the bias of DML2 decreases in K and remains approximately constant for $K \geq 10$, consistent with the explanation provided after Theorem 4.2.

6.2.2 MSE

Figure 2 presents MSE results for DML1 and DML2 across several values of K for two econometric models: ATT-DID in panel (a) and LATE in panel (b). Panel (a) shows that DML1 and DML2 perform similarly in terms of MSE, consistent with Theorem 3.1 since $\Lambda = 0$ in this case. Panel (b) shows that the MSE of DML1 increases approximately quadratically in K. This finding aligns with the expressions in Remark 4.3 for the oracle version of DML1. Additional simulation results in Figure D.3 (Appendix D) show that the DML1 estimator and its oracle version exhibit similar MSE values. Importantly, in both panels, the MSE of DML2 decreases in K and remains approximately constant for $K \geq 10$, consistent with our explanation provided after Theorem 4.2.



(a) ATT-DID ($\Lambda = 0$)

(b) LATE ($\Lambda \neq 0$)

Figure 3: Coverage probability of 95%-confidence intervals based on DML1 and DML2 for ATT-DID and LATE. Sample size $n = 3,000$ and 5,000 simulations.

6.2.3 Coverage probability

Figure 3 presents coverage probability results for 95% confidence intervals based on DML1 and DML2 across several values of K for two econometric models: ATT-DID in panel (a) and LATE in panel (b). Panel (a) shows that both DML1 and DML2 confidence intervals have similar coverage probabilities, while panel (b) shows that the coverage distortion of the DML1-based confidence interval increases in K. Corollaries 3.1 and 3.2 explain the finite-sample behavior observed in both panels.

Remark 6.1. Figures D.5 and D.6 in Appendix D report results for the ATT-DID and the LATE, respectively, for different choices of bandwidths. They show that the bias and MSE are sensitive to the choice of bandwidth, and that non-monotonic behavior of the bias can occur. □

7 Concluding remarks

This paper studies the properties of debiased machine learning (DML) estimators under a novel asymptotic framework. DML is an estimation method suited to economic models

in which the parameter of interest depends on unknown nuisance functions that must be estimated. In practice, two versions of DML—introduced by Chernozhukov et al. (2018)—can be used, that is, DML1 and DML2. Both versions randomly divide data into K equal-sized folds for estimating the nuisance function, but they differ in how these estimates are combined to estimate the parameters of interest. In this paper, we consider an asymptotic framework in which K can increase to infinity as n diverges to infinity.

This paper makes several contributions within this new framework. First, it shows that DML2 asymptotically outperforms DML1 in terms of bias, mean squared error, and inference. Additionally, it characterizes the first-order asymptotic difference between DML1 and DML2 using a discrepancy measure, Λ, which can be calculated for many econometric models. Second, it provides conditions under which all DML2 estimators, regardless of K, are asymptotically valid and share the same limiting distribution. To differentiate among them, we derive a second-order asymptotic approximations that lead to the following final contribution: setting $K = n$ for DML2 implementation is asymptotically optimal in terms of second-order asymptotic bias and MSE within the class of DML2 estimators under the conditions we provide.

A Proof of Main Results

We rely on the next two lemmas:

Lemma A.1. *Let Assumptions 3.1 and 3.2 hold and let K_n be such $K_n = O(\sqrt{n})$ and $K_n \leq n$.*

(a) Then, equation (3.3) holds.

(b) Then, equation (3.5) holds.

Lemma A.2. *Let Assumptions 3.1 (a)–(c), 3.2, and 3.3 hold and let K_n be such that $K_n \leq n$. Then, equation (3.5) holds.*

A.1 Proof of Theorems 3.1 and 3.2

Proof of Theorem 3.1. First, note that

$$\sqrt{n}\left(\hat{\theta}^{(j)}_{n,K_n} - \hat{\theta}^{*,(j)}_{n,K_n}\right) \xrightarrow{p} 0 \ ,$$

as $n \to \infty$ for $j = 1, 2$ due to Lemma A.1 in Appendix B. Second, $\hat{\theta}^{*,(2)}_{n,K_n}$ and $\hat{\theta}^{*}_{n}$ are the same to conclude that $\sqrt{n}(\hat{\theta}^{*,(2)}_{n,K_n} - \theta_0) \xrightarrow{d} N(0, \Sigma)$ by standard arguments. Finally, Lemma 3.1 in

Section 3.1 demonstrate that $\sqrt{n}(\hat{\theta}^{*,(1)}_{n,K_n} - \theta_0) \xrightarrow{d} N(c\Lambda, \Sigma)$. □

Proof of Theorem 3.2. By Lemma A.2, $\sqrt{n}\left(\hat{\theta}^{(2)}_{n,K_n} - \hat{\theta}^{*,(2)}_{n,K_n}\right) \xrightarrow{p} 0$, for $K_n = O(n)$, which is sufficient to conclude the theorem since $\sqrt{n}(\hat{\theta}^{*,(2)}_{n,K_n} - \theta_0) \xrightarrow{d} N(0, \Sigma)$. □

A.2 Proof of Lemma A.1

Proof. We use notation and auxiliary results presented in Appendix B.
Proof of part (a): We write

$$\sqrt{n}\left(\hat{\theta}^{(1)}_{n,K_n} - \hat{\theta}^{*,(1)}_{n,K_n}\right) = A + B$$

where

$$A = K_n^{-1/2} \sum_{k=1}^{K_n} \left(\mathbb{I}_d + n_k^{-1/2}\hat{b}_k\right)^{-1} (\hat{a}_k - a_k)$$
$$B = K_n^{-1/2} \sum_{k=1}^{K_n} \left\{\left(\mathbb{I}_d + n_k^{-1/2}\hat{b}_k\right)^{-1} - \left(\mathbb{I}_d + n_k^{-1/2}b_k\right)^{-1}\right\} a_k$$

and $\hat{a}_k$, $\hat{b}_k$, a_k, and b_k are defined in Appendix B.

We obtain

$$||A|| \leq ||I_1|| + \max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2}\hat{b}_k)^{-1}\right|\right| \times I_2$$

by using the identity (B.1) presented in Appendix B and the triangle inequality, and

$$||B|| \leq \max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2}b_k)^{-1}\right|\right| \times \max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2}\hat{b}_k)^{-1}\right|\right| \times I_3 \ ,$$

by using the triangle inequality and the inequality (B.2) presented in Appendix B, where

$$I_1 = K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k \tag{A.1}$$
$$I_2 = n^{-1/2} \sum_{k=1}^{K_n} \left|\left|\hat{b}_k - b_k\right|\right| \times ||\hat{a}_k - a_k||$$
$$I_3 = n^{-1/2} \sum_{k=1}^{K_n} ||b_k|| \times ||\hat{a}_k - a_k||$$
$$I_4 = n^{-1/2} \sum_{k=1}^{K_n} \left|\left|\hat{b}_k - b_k\right|\right| \times ||a_k||$$

We next show that $I_j = o_p(1)$ for $j = 1, 2, 3$, which is sufficient to complete the proof of part (a) for DML1 since Lemma B.2 guarantees that both $\max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2} b_k)^{-1}\right|\right|$ and $\max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2}\hat{b}_k)^{-1}\right|\right|$ are $O_p(1)$ when $K_n = O(n^{1/2})$.

Claim 1: $I_1 = o_p(1)$. We use Taylor expansion to write $I_1 = I_{1,1} + I_{1,2}$, where

$$I_{1,1} = n^{-1/2} \sum_{k=1}^{K_n} \sum_{i\in\mathcal{I}_k} D_\eta m_i[\hat{\eta}_i - \eta_i] \tag{A.2}$$

$$I_{1,2} = n^{-1/2} \sum_{k=1}^{K_n} \sum_{i\in\mathcal{I}_k} \frac{1}{2} D_\eta^2 \tilde{m}_i[\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \ . \tag{A.3}$$

By the Law of Iterated Expectations and part (c) of Assumption 3.1, $E[I_{1,1}] = 0$. Let e_j be the j-th column of the identity matrix $\mathbb{I}_d$. To conclude that $I_{1,1} = o_p(1)$, it is sufficient to show $E[(e_j^\top I_{1,1})^2] \to 0$. To see this, consider the following derivations,

$$\begin{aligned}
E[(e_j^\top I_{1,1})^2] &\overset{(1)}{\leq} n^{-1} K_n \sum_{k=1}^{K_n} E\left[\left(\sum_{i\in\mathcal{I}_k} e_j^\top \left(D_\eta m_i[\hat{\eta}_i - \eta_i]\right)\right)^2\right] \\
&\overset{(2)}{=} n^{-1} K_n \sum_{k=1}^{K_n} E\left[\sum_{i\in\mathcal{I}_k} \left(e_j^\top \left(D_\eta m_i[\hat{\eta}_i - \eta_i]\right)\right)^2\right] \\
&\overset{(3)}{\leq} C(n^{-1/2} K_n) n^{-1/2} \sum_{k=1}^{K_n} \sum_{i\in\mathcal{I}_k} E\left[||\hat{\eta}_k(X_i) - \eta_0(X_i)||^2\right] \\
&\overset{(4)}{=} O(1) \times o(1)
\end{aligned}$$

where (1) holds by Jensen's inequality, (2) holds because $\{e_j^\top \left(D_\eta m_i[\hat{\eta}_i - \eta_i]\right) : i \in \mathcal{I}_k\}$ are uncorrelated random variables, (3) holds by Lemma B.1, and (4) holds since $K_n = O(n^{1/2})$ and by Assumption 3.2.

The next derivations shows that $I_{1,2} = o_p(1)$,

$$\begin{aligned}
E[||I_{1,2}||] &\overset{(1)}{\leq} C n^{-1/2} \sum_{k=1}^{K_n} \sum_{i\in\mathcal{I}_k} E\left[||\hat{\eta}_k(X_i) - \eta_0(X_i)||^2\right] \\
&\overset{(2)}{=} o(1)
\end{aligned} \tag{A.4}$$

where (1) holds by the triangle inequality and Lemma B.1, and (2) holds by Assumption 3.2.

Claim 2: $I_2 = o_p(1)$. We first use the Taylor expansion to write

$$\hat{a}_k - a_k = n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta m_i[\hat{\eta}_i - \eta_i] + \frac{1}{2} D_\eta^2 \tilde{m}_i[\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i]$$

and

$$\hat{b}_k - b_k = n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta \psi_i^a[\hat{\eta}_i - \eta_i] + \frac{1}{2} D_\eta^2 \tilde{\psi^a}_i[\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \ .$$

Let $\mathcal{D}m_n$ and $\mathcal{D}\psi_n^a$ be as in Appendix B. Then,

$$\begin{aligned}
I_2 &\overset{(1)}{\leq} (n^{-1/2} K_n) \times \mathcal{D}m_n \times \mathcal{D}\psi_n^a + C\left(\mathcal{D}m_n + \mathcal{D}\psi_n^a\right) \times n^{-1/2} \sum_{k=1}^{K_n} \left(n_k^{-1/2} \sum_{i \in \mathcal{I}_k} ||\hat{\eta}_i - \eta_i||^2 \right) \\
&\quad + C^2 n^{-1/2} \sum_{k=1}^{K_n} \left(n_k^{-1/2} \sum_{i \in \mathcal{I}_k} ||\hat{\eta}_i - \eta_i||^2 \right)^2 \\
&\overset{(2)}{\leq} O(1) o_p(1) + \left(K_n^{1/2} n^{-1/2}\right) \times o_p(1) + O(1) \times \left(n^{-1/2} \sum_{i=1}^{n} ||\hat{\eta}_i - \eta_i||^2 \right)^2 \\
&\overset{(3)}{=} o_p(1) \ ,
\end{aligned}$$

where (1) holds by the triangle inequality and Lemma B.1, (2) and (3) hold by Lemmas B.2 and B.3 and since $K_n = O(n^{1/2})$. This completes proof of claim 2.

Claim 3: $I_3 = o_p(1)$. As in the proof of Claim 2, we use the Taylor expansion and $\mathcal{D}m_n$ defined in Appendix B to obtain,

$$\begin{aligned}
I_3 &\overset{(1)}{\leq} (\mathcal{D}m_n) \times n^{-1/2} \sum_{k=1}^{K_n} ||b_k|| + C n^{-1/2} \sum_{k=1}^{K_n} ||b_k|| \times \left(n_k^{-1/2} \sum_{i \in \mathcal{I}_k} ||\hat{\eta}_i - \eta_i||^2 \right) \\
&\overset{(2)}{\leq} o_p(1) \times \left(n^{-1/2} \sum_{k=1}^{K_n} ||b_k|| \right) \\
&\quad + C n^{-1} K_n^{1/2} \left(\sum_{k=1}^{K_n} ||b_k||^2 \right)^{1/2} \times \left(\sum_{k=1}^{K_n} \left(\sum_{i \in \mathcal{I}_k} ||\hat{\eta}_i - \eta_i||^2 \right)^2 \right)^{1/2} \\
&\overset{(3)}{\leq} o_p(1) \times O_p(1) + n^{-1} K_n \times O_p(1) \times \left(\sum_{k=1}^{K_n} \sum_{i \in \mathcal{I}_k} ||\hat{\eta}_i - \eta_i||^2 \right) \\
&\overset{(4)}{=} o_p(1) \ ,
\end{aligned}$$

where (1) holds by the triangle inequality and Lemma B.1, (2) holds by Lemma B.2 and the Cauchy-Schwarz inequality, (3) holds by part (b) of Assumption 3.1, using the definition of

b_k, and since $K_n = O(n^{1/2})$, and (4) holds by Assumption 3.2 and since $K_n = O(n^{1/2})$.
Claim 4: $I_4 = o_p(1)$. The proof is similar to Claim 4 but using $\mathcal{D}\psi_n^a$ instead of $\mathcal{D}m_n$; therefore, omitted.
Proof of part (b): We write

$$\sqrt{n}\left(\hat{\theta}^{(2)}_{n,K_n} - \hat{\theta}^{*,(2)}_{n,K_n}\right) = A + B \ ,$$

where

$$A = \left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1} \left(K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k\right)$$
$$B = \left\{\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1} - \left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} b_k\right)^{-1}\right\} \left(K_n^{-1/2} \sum_{k=1}^{K_n} a_k\right)$$

and $\hat{a}_k$, $\hat{b}_k$, a_k, and b_k are defined in Appendix B.

A is $o_p(1)$ due to two results. First, $\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1}$ is $O_p(1)$ by Lemma B.3. Second, $K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k$ is $o_p(1)$ by claim 1 in the proof of part (a).

To show that B is $o_p(1)$, we consider the following derivations

$$\begin{aligned}
||B|| &\overset{(1)}{\leq} \left|\left|\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1}\right|\right| \times \left|\left|\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} b_k\right)^{-1}\right|\right| \\
&\quad \times \left|\left| n^{-1/2} K_n^{-1/2} \sum_{k=1}^{K_n} \hat{b}_k - b_k \right|\right| \times \left|\left| K_n^{-1/2} \sum_{k=1}^{K_n} a_k \right|\right| \\
&\overset{(2)}{\leq} O_p(1) \times n^{-1/2} \left|\left| K_n^{-1/2} \sum_{k=1}^{K_n} \hat{b}_k - b_k \right|\right| \\
&\overset{(3)}{=} o_p(1) \ ,
\end{aligned}$$

where (1) holds by the inequality (B.2) in Appendix B, (2) holds by Lemma B.3 and the Central Limit Theorem, and (3) holds since $\left|\left| K_n^{-1/2} \sum_{k=1}^{K_n} \hat{b}_k - b_k \right|\right|$ is $o_p(1)$ due to the same arguments we used to prove that $I_1 = o_p(1)$ in claim 1 in the proof of part (a) for DML1. □

A.3 Proof of Lemma A.2

Proof. The proof of part (b) in Lemma A.1 relies on Lemma B.3 and $I_1 = o_p(1)$, where I_1 is defined in (A.1). Lemma B.3 holds for $K_n = O(n)$, but the proof of $I_1 = o_p(1)$ relies

on $K_n = O(n^{1/2})$. Therefore, the validity of the previous proof does not apply to the case $K_n = O(n)$. To adapt the proof of part (b) in Lemma B.3, we show that $I_1 = o_p(1)$ also holds when $K_n = O(n)$, provided we add Assumption 3.3.

Recall that $I_1 = I_{1,1} + I_{1,2}$, where $I_{1,1}$ and $I_{1,2}$ are defined in (A.2) and (A.3), respectively. Note that the proof of $I_{1,2} = o_p(1)$ also applies when $K_n = O(n)$; see derivations in (A.4). Therefore, it is sufficient to show that $I_{1,1} = o_p(1)$. Since $E[I_{1,1}] = 0$, it is sufficient to show that $E[(e_j^\top I_{1,1})^2] = o(1)$, where e_j is the j-th column of $\mathbb{I}_d$. Consider the following derivations,

$$\begin{aligned}
E[(e_j^\top I_{1,1})^2] &= E\left[\left(n^{-1/2}\sum_{k=1}^{K_n}\sum_{i\in\mathcal{I}_k} e_j^\top(D_\eta m_i[\hat{\eta}_i - \eta_i])\right)^2\right] \\
&= n^{-1}\sum_{k_1=1}^{K_n}\sum_{k_2=1}^{K_n}\sum_{i_1\in\mathcal{I}_{k_1}}\sum_{i_2\in\mathcal{I}_{k_2}} E[e_j^\top(D_\eta m_{i_1}[\hat{\eta}_{i_1} - \eta_{i_1}])e_j^\top(D_\eta m_{i_2}[\hat{\eta}_{i_2} - \eta_{i_2}])] \\
&= I_{1,1,1} + I_{1,1,2} + I_{1,1,3}
\end{aligned}$$

where

$$\begin{aligned}
I_{1,1,1} &= n^{-1}\sum_{k=1}^{K_n}\sum_{i\in\mathcal{I}_k} E\left[\left(e_j^\top(D_\eta m_i[\hat{\eta}_i - \eta_i])\right)^2\right] \\
I_{1,1,2} &= n^{-1}\sum_{k=1}^{K_n}\sum_{i_1,i_2\in\mathcal{I}_k} E[e_j^\top(D_\eta m_{i_1}[\hat{\eta}_{i_1} - \eta_{i_1}])e_j^\top(D_\eta m_{i_2}[\hat{\eta}_{i_2} - \eta_{i_2}])]I\{i_1 \neq i_2\} \\
I_{1,1,3} &= n^{-1}\sum_{k_1,k_2=1}^{K_n}\sum_{i_1\in\mathcal{I}_{k_1}}\sum_{i_2\in\mathcal{I}_{k_2}} E[e_j^\top(D_\eta m_{i_1}[\hat{\eta}_{i_1} - \eta_{i_1}])e_j^\top(D_\eta m_{i_2}[\hat{\eta}_{i_2} - \eta_{i_2}])]I\{k_1 \neq k_2\}
\end{aligned}$$

Note that $I_{1,1,1} = o(1)$ and $I_{1,1,2} = 0$. The former holds by Assumption 3.2 and $|I_{1,1,1}| \leq Cn^{-1}\sum_{i=1}^n E\left[||\hat{\eta}_i - \eta_i||^2\right]$, while the latter by part (c) of Assumption 3.1 and the Law of Iterated Expectations.

We now show that $I_{1,1,3} = o(1)$ using Assumption 3.3. We proceed in three steps. First, for $i_1 \in \mathcal{I}_{k_1}$, $i_2 \in \mathcal{I}_{k_2}$, and $k_1 \neq k_2$, let $\hat{\eta}_{i_1}^{i_2} = \hat{\eta}_{k_1}^{i_2}(X_{i_1})$ and $\hat{\eta}_{i_2}^{i_1} = \hat{\eta}_{k_2}^{i_1}(X_{i_2})$. We have

$$E[e_j^\top(D_\eta m_{i_1}[\hat{\eta}_{i_1}^{i_2} - \eta_{i_1}])e_j^\top(D_\eta m_{i_2}[\hat{\eta}_{i_2} - \eta_{i_2}])] = 0\ ,$$

which holds by the Law of Iterated Expectations (conditional on X_{i_2} and $\{W_i : 1 \leq i \leq n, i \neq i_2\}$), part (c) of Assumption 3.1, and the definition of $\hat{\eta}_{k_1}^{i_2}(X_{i_1})$. Second, we use that $D_\eta m_{i_1}$ is a linear operator (i.e., $D_\eta m_{i_1}[\hat{\eta}_{i_1} - \eta_{i_1}] = D_\eta m_{i_1}[\hat{\eta}_{i_1} - \hat{\eta}_{i_1}^{i_2}] + D_\eta m_{i_1}[\hat{\eta}_{i_1}^{i_2} - \eta_{i_1}]$) and

the previous step to write

$$I_{1,1,3} = n^{-1} \sum_{k_1,k_2=1}^{K_n} \sum_{i_1 \in \mathcal{I}_{k_1}} \sum_{i_2 \in \mathcal{I}_{k_2}} E[e_j^\top (D_\eta m_{i_1}[\hat{\eta}_{i_1} - \hat{\eta}_{i_1}^{i_2}]) e_j^\top (D_\eta m_{i_2}[\hat{\eta}_{i_2} - \hat{\eta}_{i_2}^{i_1}])] I\{k_1 \neq k_2\} \ .$$

Finally, we use the previous step, the Cauchy-Schwarz inequality, and Lemma B.1 to obtain

$$\begin{aligned} I_{1,1,3} &\leq C n^{-1} \sum_{k_1,k_2=1}^{K_n} \sum_{i_1 \in \mathcal{I}_{k_1}} \sum_{i_2 \in \mathcal{I}_{k_2}} E\left[\left|\left|\hat{\eta}_{i_1} - \hat{\eta}_{i_1}^{i_2}\right|\right|^2\right]^{1/2} E\left[\left|\left|\hat{\eta}_{i_2} - \hat{\eta}_{i_1}^{i_2}\right|\right|^2\right]^{1/2} \\ &\overset{(1)}{=} o(1) \ , \end{aligned}$$

where (1) holds by Assumption 3.3. This completes the proof of $I_1 = o_p(1)$. □

A.4 Proof of Lemma 3.1

Proof. We use the definition of $\hat{\theta}_{n,K_n}^{*,(1)}$ to write

$$\sqrt{n}\left(\hat{\theta}_{n,K_n}^{*,(1)} - \theta_0\right) = K_n^{-1/2} \sum_{k=1}^{K_n} \left(\mathbb{I}_d + n_k^{-1/2} b_k\right)^{-1} a_k \ ,$$

where

$$\begin{aligned} a_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} G^{-1} m_i \\ b_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} \left(G^{-1} \psi_i^a - \mathbb{I}_d\right) \ . \end{aligned}$$

We now use the identity

$$(\mathbb{I}_k + n_k^{-1/2} b_k)^{-1} a_k = a_k - n_k^{-1/2} b_k a_k + n_k^{-1} (\mathbb{I}_d + n_k^{-1/2} b_k)^{-1} b_k^2 a_k$$

to write

$$\sqrt{n}\left(\hat{\theta}_{n,K_n}^{*,(1)} - \theta_0\right) - (K_n/\sqrt{n})\Lambda = I_1 - I_2 + I_3$$

where

$$I_1 = K_n^{-1/2} \sum_{k=1}^{K_n} a_k = n^{-1/2} \sum_{i=1}^{n} G^{-1} m_i$$

$$I_2 = K_n^{-1/2} \sum_{k=1}^{K_n} n_k^{-1/2} b_k a_k + (K_n/\sqrt{n})\Lambda$$
$$I_3 = K_n^{-1/2} \sum_{k=1}^{K_n} n_k^{-1} (\mathbb{I}_d + n_k^{-1/2} b_k)^{-1} b_k^2 a_k$$

Claims 1 and 2 below show that $I_2 = o_p(1)$ and $I_3 = o_p(1)$, which is sufficient to complete the proof of this lemma since $I_1 \xrightarrow{d} N(0, \Sigma)$ by the Central Limit Theorem.

Claim 1: $I_2 = o_p(1)$. To show this, we first note that $E[I_2] = 0$ since $E[b_k a_k] = -\Lambda$. It is sufficient to show that $E[||I_2||^2] \to 0$. Algebra shows

$$\begin{aligned}
E[||I_2||^2] &\overset{(1)}{=} E\left[\left|\left|n^{-1/2} \sum_{k=1}^{K_n} (b_k a_k - E[b_k a_k])\right|\right|^2\right] \\
&\overset{(2)}{=} n^{-1} \sum_{k=1}^{K_n} E\left[||(b_k a_k - E[b_k a_k])||^2\right] \\
&\overset{(3)}{\leq} n^{-1} K_n E\left[||b_k a_k||^2\right] \\
&\overset{(4)}{\leq} n^{-1} K_n E\left[||b_k||^4\right]^{1/2} E\left[||a_k||^4\right]^{1/2} \\
&\overset{(5)}{=} n^{-1} K_n \times O(1) \times O(1) \ ,
\end{aligned}$$

where (1) holds since $E[b_k a_k] = -\Lambda$, (2) and (3) hold because $\{(b_k a_k - E[b_k a_k]) : 1 \leq k \leq K_n\}$ are i.i.d. zero mean random vectors, (4) holds by CS inequality, and (5) holds by part (b) Assumption 3.1 and using the definition of a_k and b_k. Therefore, $E[||I_2||^2] = O(n^{-1/2})$ since $K_n = O(n^{1/2})$.

Claim 2: $I_3 = o_p(1)$. To show this, first note that

$$||I_3|| \leq \max_{k=1,\ldots,K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2} b_k)^{-1}\right|\right| \times K_n^{-1/2} \sum_{k=1}^{K_n} n_k^{-1} ||b_k^2 a_k|| \ ,$$

where $\max_{k=1,\ldots,K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2} b_k)^{-1}\right|\right| = O_p(1)$ due to Lemma B.2. Therefore, it is sufficient to show that $K_n^{-1/2} \sum_{k=1}^{K_n} n_k^{-1} ||b_k^2 a_k|| = o_p(1)$, which holds by the following derivations

$$\begin{aligned}
E[K_n^{-1/2} \sum_{k=1}^{K_n} n_k^{-1} ||b_k^2 a_k||] &\overset{(1)}{\leq} K_n^{3/2} n^{-1} E[||b_k||^4]^{1/2} E[||a_k||^2]^{1/2} \\
&\overset{(2)}{\leq} K_n^{3/2} n^{-1} \times O(1) \times O(1) \\
&\overset{(3)}{=} O(n^{-1/4}) \ ,
\end{aligned}$$

where (1) holds because $\{b_k^2 a_k : 1 \leq k \leq K_n\}$ are i.i.d. random vectors and Cauchy-Schwarz inequality, (2) holds by part (b) of Assumption 3.1 and using the definition of a_k and b_k, and (3) holds since $K_n = O(n^{1/2})$. This completes the proof of claim 2. □

B Auxiliary Results

We use the following notation in the proofs of the main results in Appendix A.

$$\begin{aligned}
\hat{a}_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} G^{-1} \hat{m}_i \\
\hat{b}_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} \left(G^{-1} \hat{\psi}_i^a - \mathbb{I}_d \right) \\
a_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} G^{-1} m_i \\
b_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} \left(G^{-1} \psi_i^a - \mathbb{I}_d \right) \\
\mathcal{D}m_n &= \max_{1 \leq k \leq K_n} \left|\left| n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta m_i [\hat{\eta}_i - \eta_i] \right|\right| \\
\mathcal{D}\psi_n^a &= \max_{1 \leq k \leq K_n} \left|\left| n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta \psi_i^a [\hat{\eta}_i - \eta_i] \right|\right|
\end{aligned}$$

We also use the following identity

$$(\mathbb{I}_d + \mathbb{M})^{-1} = \mathbb{I}_d - (\mathbb{I}_d + \mathbb{M})^{-1} \mathbb{M} \tag{B.1}$$

and inequality

$$\left|\left|(\mathbb{I}_d + \mathbb{M}_1)^{-1} - (\mathbb{I}_d + \mathbb{M}_2)^{-1}\right|\right| \leq \left|\left|(\mathbb{I}_d + \mathbb{M}_1)^{-1}\right|\right| \cdot ||\mathbb{M}_1 - \mathbb{M}_2|| \cdot \left|\left|(\mathbb{I}_d + \mathbb{M}_2)^{-1}\right|\right| \tag{B.2}$$

The next lemmas are used in the proof of the main results in Appendix A.

Lemma B.1. *Let Assumption 3.1 holds. Then, there exists a constant $C > 0$ such that*

1. $E[(e_j^\top (D_\eta m_{i_1} [\hat{\eta}_{i_1} - \eta_{i_1}]))] \leq C E[||\hat{\eta}_i - \eta_i||^2]$
2. $E[(e_j^\top (D_\eta m_{i_1} [\hat{\eta}_{i_1} - \hat{\eta}_{i_1}^{i_2}]))] \leq C E[\left|\left|\hat{\eta}_i - \hat{\eta}_{i_1}^{i_2}\right|\right|^2]$
3. $\frac{1}{2} \left|\left| D_\eta^2 \tilde{m}_i [\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \right|\right| \leq C ||\hat{\eta}_i - \eta_i||^2$
4. $\frac{1}{2} \left|\left| D_\eta^2 \tilde{\psi^a}_i [\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \right|\right| \leq C ||\hat{\eta}_i - \eta_i||^2$

for any $k \leq K_n$ *and* $i \in \mathcal{I}_k$.

Lemma B.2. *Let Assumptions 3.1 and 3.2 hold and let* K_n *be such that* $K_n \leq n$ *and* $K_n = O(\sqrt{n})$. *Then,*

1. $\max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2} b_k)^{-1}\right|\right| = O_p(1)$
2. $\max_{1\leq k\leq K_n} \left|\left|(\mathbb{I}_d + n_k^{-1/2} \hat{b}_k)^{-1}\right|\right| = O_p(1)$
3. $\mathcal{D}m_n = \max_{1\leq k\leq K_n} \left|\left|n_k^{-1/2} \sum_{i\in\mathcal{I}_k} D_\eta m_i[\hat{\eta}_i - \eta_i]\right|\right| = o_p(1)$
4. $\mathcal{D}\psi_n^a = \max_{1\leq k\leq K_n} \left|\left|n_k^{-1/2} \sum_{i\in\mathcal{I}_k} D_\eta \psi_i^a[\hat{\eta}_i - \eta_i]\right|\right| = o_p(1)$

Lemma B.3. *Let Assumptions 3.1 and 3.2 hold and let* K_n *be such that* $K_n \leq n$. *Then,*

1. $\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} b_k\right)^{-1} = O_p(1)$
2. $\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1} = O_p(1)$
3. $n^{-1/2} \sum_{i=1}^{n} ||\hat{\eta}_i - \eta_i||^2 = o_p(1)$

C Examples

Example C.1 (Average Treatment Effect)**.** Let $A \in \{0, 1\}$ denote a binary treatment status, $Y(a)$ denote the potential outcome under treatment $a \in \{0, 1\}$, X denote a vector of covariates, and

$$Y = AY(1) + (1 - A)Y(0)$$

denote the observed outcome. The available data is modeled by the vector $W = (Y, A, X)$. The parameter of interest is

$$\theta_0 = E[Y(1) - Y(0)] \ ,$$

which is the expectation of the treatment effect when the treatment is mandated across the entire population, also known as the ATE. A standard assumption used to identify θ_0 is the selection-on-observables assumption,

$$(Y(1), Y(0)) \perp A \mid X \ .$$

Under the selection-on-observables assumption, the ATE can be identified by a moment condition such as (2.1) using a moment function like (2.2), which is defined by

$$\psi^b(W, \eta) = \eta_1 - \eta_2 + A(Y - \eta_1)\eta_3 - (1 - A)(Y - \eta_2)\eta_4 \ ,$$

$$\psi^a(W,\eta) = 1 \ ,$$

for $\eta \in \mathbf{R}^4$, and where the nuisance parameter $\eta_0(X)$ has four components:

$$\begin{aligned}
\eta_{0,1}(X) &= E[Y \mid X, A=1] \ , \\
\eta_{0,2}(X) &= E[Y \mid X, A=0] \ , \\
\eta_{0,3}(X) &= (E[A \mid X])^{-1} \ , \\
\eta_{0,4}(X) &= (E[1-A \mid X])^{-1} \ .
\end{aligned}$$

This moment function corresponds to the augmented inverse propensity weighted (AIPW) estimator (Robins et al. (1994), Scharfstein et al. (1999)). It also appears as the efficient influence function for the ATE in Hahn (1998) and Hirano et al. (2003). □

Example C.2 (Difference-in-Differences)**.** This example considers the average treatment effect on the treated in difference-in-differences research designs with two periods and panel data, as studied in Sant'Anna and Zhao (2020). Let $A \in \{0,1\}$ denote a binary treatment status on the post-treatment period, $Y_1(a)$ denote the potential outcome on the post-treatment period under treatment status $a \in \{0,1\}$, Y_0 denote the outcome of interest in a pre-treatment period, X denote a vector of covariates, and

$$Y_1 = AY_1(1) + (1-A)Y_1(0)$$

denote the observed outcome in the post-treatment period. The available data is modeled by the vector $W = (Y_0, Y_1, A, X)$. The parameter of interest is

$$\theta_0 = E[Y_1(1) - Y_1(0) \mid A = 1] \ ,$$

which represents the treatment effect for the treated group in the post-treatment period, also known as ATT-DID. Sant'Anna and Zhao (2020) used the following conditional parallel trend assumption,

$$E[Y_1(0) - Y_0 \mid X, A=1] = E[Y_1(0) - Y_0 \mid X, A=0] \ ,$$

to identify the ATT-DID by a moment condition, such as (2.1), using a moment function like (2.2), which is defined by

$$\begin{aligned}
\psi^b(W,\eta) &= A(Y_1 - Y_0 - \eta_1) + (1-A)(1-\eta_2)(Y_1 - Y_0 - \eta_1) \ , \\
\psi^a(W,\eta) &= A \ ,
\end{aligned}$$

for $\eta \in \mathbf{R}^2$, and where the nuisance parameter $\eta_0(X)$ has two components:

$$\eta_{0,1}(X) = E[Y_1 - Y_0 \mid X, A = 0]$$
$$\eta_{0,2}(X) = (E[1 - A \mid X])^{-1} .$$

This moment function is the efficient influence function for the ATT-DID under the conditions in Sant'Anna and Zhao (2020). □

Example C.3 (Local Average Treatment Effect)**.** This example considers a framework where individuals can decide their treatment status as in Imbens and Angrist (1994) and Frölich (2007). Let $Z \in \{0, 1\}$ denote a binary instrumental variable (e.g., treatment assignment), $D(z)$ denote potential treatment decisions under the intervention $z \in \{0, 1\}$, and assume the observed treatment decision is given by

$$D = ZD(1) + (1 - Z)D(0) .$$

Let X denote a vector of covariates, $Y(d)$ denote the potential outcome under treatment decision $d \in \{0, 1\}$, and $Y = DY(1) + (1 - D)Y(0)$ denote the observed outcome. The available data is modeled by the vector $W = (Y, Z, D, X)$. The parameter of interest is

$$\theta_0 = E[Y(1) - Y(0) \mid D(1) > D(0)] ,$$

which is the expected treatment effect for the sub-population that complies with the assigned treatment, also known as LATE. A sufficient assumption for identification is the following selection-on-observables assumption,

$$(Y(1), Y(0), D(1), D(0)) \perp Z \mid X .$$

Using this assumption and similar assumptions as in Frölich (2007), Singh and Sun (2024) identified the LATE by a moment condition, such as (2.1), using a moment function like (2.2), which is defined by

$$\psi^b(W, \eta) = \eta_1 - \eta_2 + Z(Y - \eta_1)\eta_5 - (1 - Z)(Y - \eta_2)\eta_6$$
$$\psi^a(W, \eta) = \eta_3 - \eta_4 + Z(D - \eta_3)\eta_5 - (1 - Z)(D - \eta_4)\eta_6$$

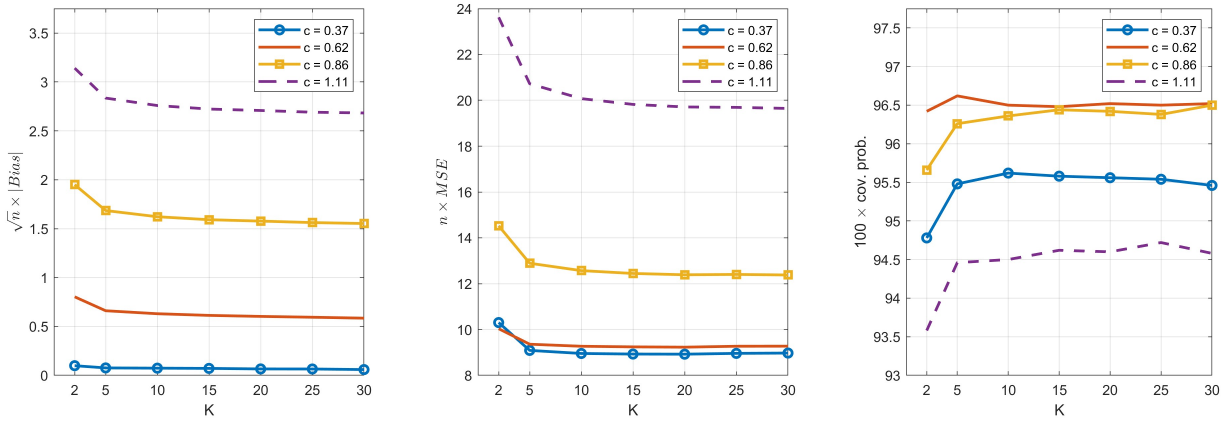
for $\eta \in \mathbf{R}^6$, and where the nuisance parameter $\eta_0(X)$ has six components:

$$\eta_{0,1}(X) = E[Y \mid X, Z = 1] ,$$

$$
\begin{aligned}
\eta_{0,2}(X) &= E[Y \mid X, Z = 0] \ , \\
\eta_{0,3}(X) &= E[D \mid X, Z = 1] \ , \\
\eta_{0,4}(X) &= E[D \mid X, Z = 0] \ , \\
\eta_{0,5}(X) &= (E[Z \mid X])^{-1} \ , \\
\eta_{0,6}(X) &= (E[1 - Z \mid X])^{-1} \ .
\end{aligned}
$$

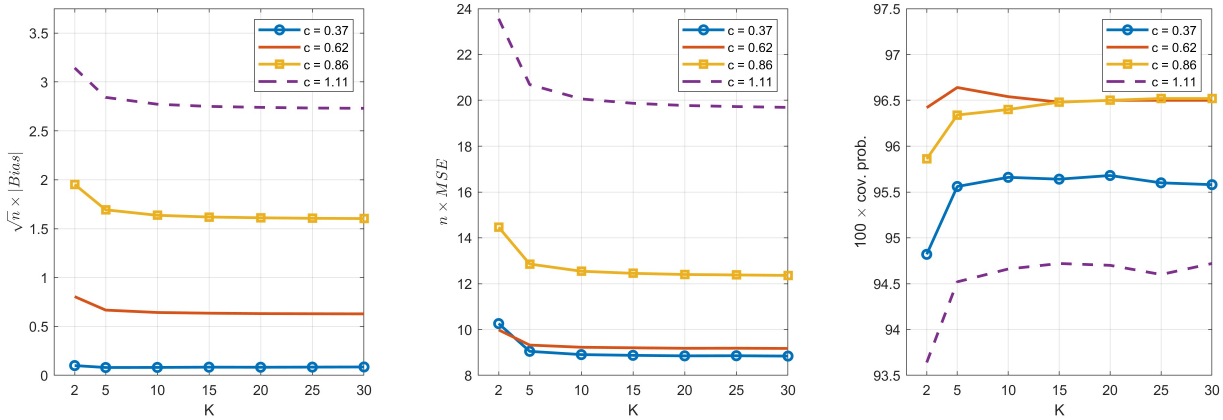
This moment function appears in Frölich (2007) as the efficient influence function for the LATE. This moment function corresponds to the estimators proposed in Tan (2006). □

D Additional Simulation Results



(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.1: Bias and MSE of estimators for the ATT-DID based on DML1 for different values of c in $h = cn_0^{-1/5}$. It uses a Second Order Gaussian Kernel Coverage probability of 95%-confidence intervals for the ATT-DID. Sample size $n = 3,000$ and 5,000 simulations.

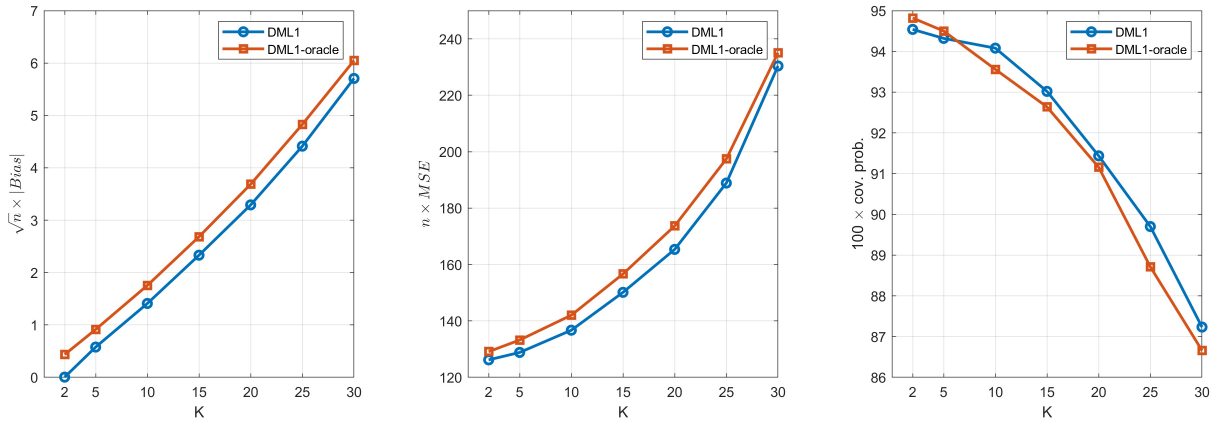


(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.2: Bias and MSE of estimators for the ATT-DID based on DML2 for different values of c in $h = cn_0^{-1/5}$. It uses a Second Order Gaussian Kernel. Coverage probability of 95%-confidence intervals for the ATT-DID. Sample size $n = 3,000$ and 5,000 simulations.

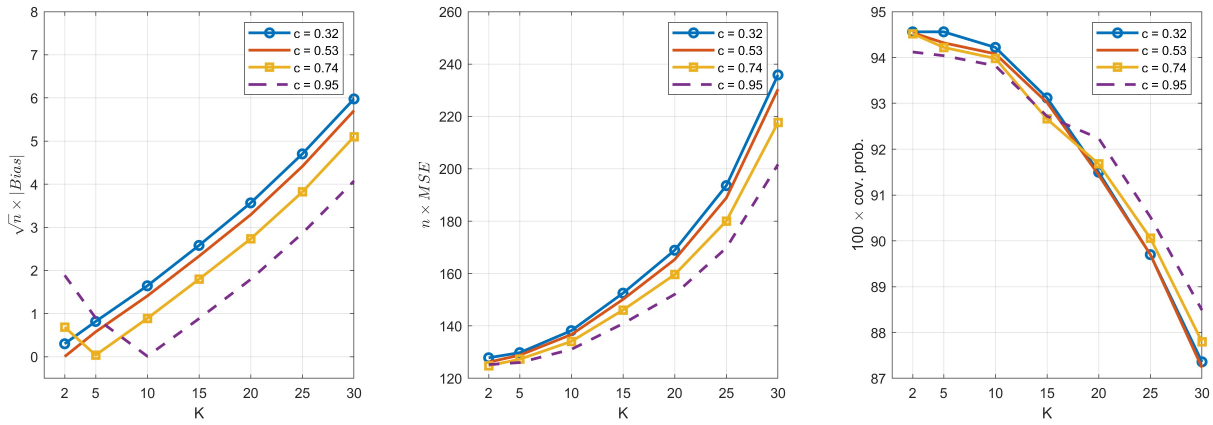
References

Ahrens, A., V. Chernozhukov, C. Hansen, D. Kozbur, M. Schaffer, and T. Wiemann (2025): "An introduction to double/debiased machine learning," *arXiv*



(a) Bias (b) MSE (c) Coverage Prob.(%)

Figure D.3: Bias and MSE of DML1 and its oracle version for the LATE. Coverage probability of 95%-confidence intervals based on DML1 and its oracle version (use true Σ) for the LATE. Discrepancy measure $\Lambda \neq 0$, sample size $n = 3,000$ and $5,000$ simulations.



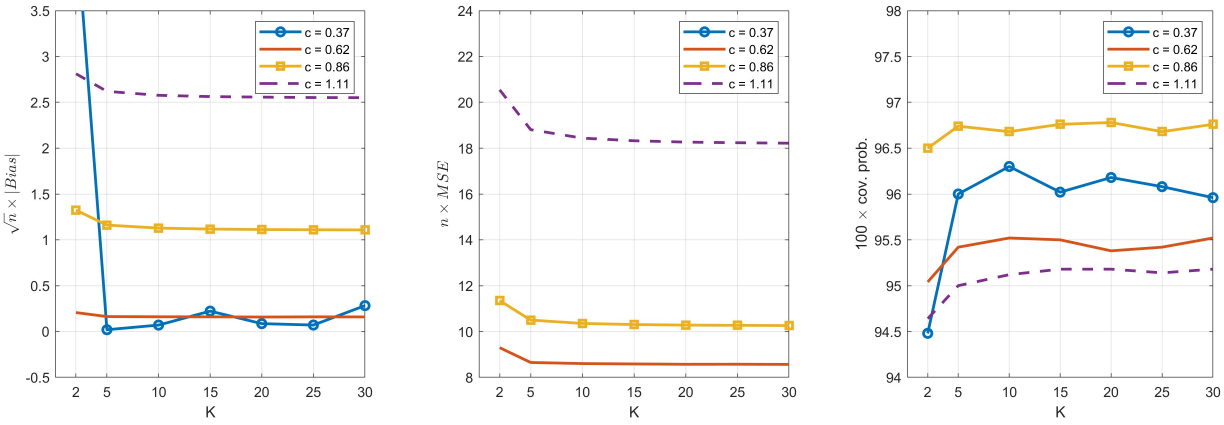
(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.4: Bias and MSE of estimators for the LATE based on DML1 for different values of c in $h = cn_0^{-1/5}$. Coverage probability of 95%-confidence intervals for the LATE . Discrepancy measure $\Lambda \neq 0$, sample size $n = 3,000$ and $5,000$.

preprint arXiv:2504.08324.

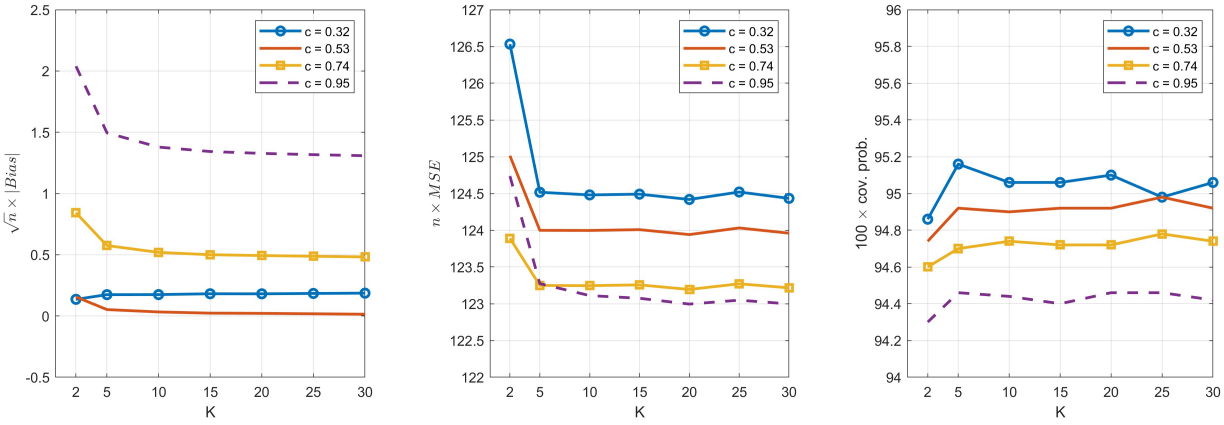
Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann (2024a): "ddml: Double/debiased machine learning in Stata," *The Stata Journal*, 24, 3–45.

——— (2024b): "Model averaging and double machine learning," *arXiv preprint arXiv:2401.01645*.



(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.5: Bias and MSE of estimators for the ATT-DID based on DML2 for different values of c in $h = cn_0^{-1/16}$. All the values of $n \times MSE$ for $c = 0.37$ are larger than 24. Coverage probability of 95%-confidence intervals for the ATT-DID. Sample size $n = 3,000$ and 5,000 simulations.



(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.6: Bias and MSE of estimators for the LATE based on DML2 for different values of c in $h = cn_0^{-1/5}$. Coverage probability of 95%-confidence intervals for the LATE. Sample size $n = 3,000$ and 5,000 simulations.

Andrews, D. W. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, 43–72.

Argañaraz, F. (2025): "Automatic Debiased Machine Learning of Structural Parameters with General Conditional Moments," *arXiv preprint arXiv:2512.08423*.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2022): "DoubleML-an

object-oriented implementation of double machine learning in python," *Journal of Machine Learning Research*, 23, 1–6.

Bach, P., M. S. Kurz, V. Chernozhukov, M. Spindler, and S. Klaassen (2024): "DoubleML: An Object-Oriented Implementation of Double Machine Learning in R," *Journal of Statistical Software*, 108, 1–56.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

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

Belloni, A., V. Chernozhukov, and L. Wang (2011): "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806.

Bickel, P. J. (1982): "On adaptive estimation," *The Annals of Statistics*, 647–671.

Bickel, P. J. and Y. Ritov (2003): "Nonparametric estimators which can be" plugged-in"," *The Annals of Statistics*, 31, 1033–1053.

Bousquet, O. and A. Elisseeff (2002): "Stability and generalization," *Journal of machine learning research*, 2, 499–526.

Bugni, F. A. and I. A. Canay (2021): "Testing continuity of a density via g-order statistics in the regression discontinuity design," *Journal of Econometrics*, 221, 138–159.

Cai, Y. (2022): "Linear Regression with Centrality Measures," *arXiv preprint arXiv:2210.10024*.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of econometrics*, 225, 200–230.

Cattaneo, M. D. and M. Jansson (2018): "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency," *Econometrica*, 86, 955–995.

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

Chen, Q., V. Syrgkanis, and M. Austern (2022): "Debiased machine learning without sample-splitting for stable estimators," *Advances in Neural Information Processing Systems*, 35, 3096–3109.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Cheng, X., A. Sánchez-Becerra, and A. J. Shephard (2023): "How to Weight in Moments Matching: A New Approach and Applications to Earnings Dynamics," *CEMMAP working paper CWP13/23*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/debiased/neyman machine learning of treatment effects," *American Economic Review*, 107, 261–265.

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

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

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

——— (2022c): "Debiased machine learning of global and local parameters using regularized Riesz representers," *The Econometrics Journal*, 25, 576–601.

Chi, C.-M., P. Vossler, Y. Fan, and J. Lv (2022): "Asymptotic properties of high-dimensional random forests," *The Annals of Statistics*, 50, 3415–3438.

Donald, S. G. and W. K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161–1191.

Escanciano, J. C. and T. Pérez-Izquierdo (2023): "Automatic Locally Robust Estimation with Generated Regressors," *arXiv preprint arXiv:2301.10643*.

Escanciano, J. C. and J. R. Terschuur (2023): "Machine Learning Inference on Inequality of Opportunity," .

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

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

——— (2025): "Deep learning for individual heterogeneity: An automatic inference framework," *arXiv preprint arXiv:2010.14694*.

Fava, B. (2024): "Predicting the Distribution of Treatment Effects: A Covariate-Adjustment Approach," *arXiv preprint arXiv:2407.14635*.

Frölich, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35–75.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012): "Inverse probability tilting for moment condition models with missing data," *The Review of Economic Studies*, 79, 1053–1079.

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

Hahn, J. and G. Ridder (2013): "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 81, 315–340.

Hardt, M., B. Recht, and Y. Singer (2016): "Train faster, generalize better: Stability of stochastic gradient descent," in *International conference on machine learning*, PMLR, 1225–1234.

Hastie, T., R. Tibshirani, J. Friedman, et al. (2009): "The elements of statistical learning," .

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

Hong, H. and D. Nekipelov (2010): "Semiparametric efficiency in nonlinear LATE models," *Quantitative Economics*, 1, 279–304.

Ichimura, H. and W. K. Newey (2022): "The influence function of semiparametric estimators," *Quantitative Economics*, 13, 29–61.

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

Ji, W., L. Lei, and A. Spector (2023): "Model-agnostic covariate-assisted inference on partially identified causal effects," *arXiv preprint arXiv:2310.08115*.

Jin, J. and V. Syrgkanis (2024): "Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation," *arXiv preprint arXiv:2402.14264*.

Kennedy, E. H., S. Balakrishnan, J. M. Robins, and L. Wasserman (2024): "Minimax rates for heterogeneous causal effect estimation," *The Annals of Statistics*, 52, 793–816.

Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica: Journal of the Econometric Society*, 1079–1112.

Liu, Y. and F. Molinari (2025): "Inference for an Algorithmic Fairness-Accuracy Frontier," *arXiv preprint arXiv:2402.08879*.

Liu, Y., F. Molinari, and A. Velez (2026): "Identification and Inference for Algorithmic Frontiers with Selective Labels," *Work in progress*.

Newey, W. K. (1990): "Efficient instrumental variables estimation of nonlinear models," *Econometrica: Journal of the Econometric Society*, 809–837.

——— (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79, 147–168.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K. and J. R. Robins (2018): "Cross-fitting and fast remainder rates for semiparametric estimation," *arXiv preprint arXiv:1801.09138*.

Newey, W. K. and R. J. Smith (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

Noack, C., T. Olma, and C. Rothe (2024): "Flexible covariate adjustments in regression discontinuity designs," *arXiv preprint arXiv:2107.07942*.

Park, G. (2024): "Debiased Machine Learning when Nuisance Parameters Appear in Indicator Functions," *arXiv preprint arXiv:2403.15934*.

Rafi, A. (2023): "Efficient semiparametric estimation of average treatment effects under covariate adaptive randomization," *arXiv preprint arXiv:2305.08340*.

Robins, J. M. and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

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

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, 931–954.

Rothe, C. and S. Firpo (2019): "Properties of doubly robust estimators when nuisance functions are estimated nonparametrically," *Econometric Theory*, 35, 1048–1087.

Rothenberg, T. J. (1984): "Approximating the distributions of econometric estimators and test statistics," *Handbook of econometrics*, 2, 881–935.

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

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, 94, 1096–1120.

Schmidt-Hieber, J. (2020): "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, 48, 1875 – 1897.

Semenova, V. (2023a): "Adaptive estimation of intersection bounds: a classification approach," *arXiv preprint arXiv:2303.00982*.

——— (2023b): "Debiased machine learning of set-identified linear models," *Journal of Econometrics*, 235, 1725–1746.

Semenova, V. and V. Chernozhukov (2021): "Debiased machine learning of conditional average treatment effects and other causal functions," *The Econometrics Journal*, 24, 264–289.

Singh, R. and L. Sun (2024): "Double robustness for complier parameters and a semiparametric test for complier characteristics," *The Econometrics Journal*, 27, 1–20.

Tan, Z. (2006): "Regression and weighting methods for causal inference using instrumental variables," *Journal of the American Statistical Association*, 101, 1607–1618.

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

Van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," .