On the Asymptotic Properties of Debiased Machine Learning Estimators *

Amilcar Velez
Department of Economics
Northwestern University
amilcare@u.northwestern.edu

This version: November 19, 2024.
Newest version here.

Abstract

This paper studies the properties of debiased machine learning (DML) estimators under a novel asymptotic framework, offering insights for improving estimator performance in applications. DML is an estimation method suited to economic models in which the parameter of interest depends on unknown nuisance functions that must be estimated. It requires weaker conditions than previous methods while still ensuring standard asymptotic properties. Existing theoretical results do not distinguish between the two alternative versions of DML estimators, namely, DML1 and DML2. Under a new asymptotic framework, this paper demonstrates that DML2 asymptotically dominates DML1 in terms of bias and mean squared error, formalizing a previous conjecture based on the simulation results of their relative performance. Additionally, this paper provides guidance for improving DML2 performance in applications.

*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 am also thankful to Eric Auerbach, Federico Crippa, Igal Hendel, Diego Huerta, Danil Fedchenko, Chuck Manski, Giorgio Primiceri, Sebastian Sardon, Chris Walker, and Thomas Wiemann for valuable comments and suggestions. Financial support from the Robert Eisner Memorial Fellowship and the Dissertation Year Fellowship is gratefully acknowledged. Any and all errors are my own.

1 Introduction

Debiased machine learning (DML) has become a popular method for estimating parameters in economic models. DML is particularly suited to cases where the parameter of interest depends on unknown nuisance functions that require estimation (Chernozhukov et al., 2018). DML offers standard asymptotic properties (e.g., asymptotic normality and parametric convergence rate) under milder conditions compared to previous methods (e.g., Newey (1994), Andrews (1994), Newey and McFadden (1994)). In practice, two versions of DML —introduced by Chernozhukov et al. (2018)—can be used, DML1 and DML2. Both versions randomly divide the data into K equal-sized folds (samples) to estimate the nuisance function, but they differ in how these estimates are used to construct an estimator for the parameter of interest. While DML2 is believed to perform better than DML1 based on the simulation results of their relative performance, DML1 and DML2 yield estimators with the same asymptotic distribution when K remains fixed as the sample size n diverges to infinity. This paper studies the properties of DML1 and DML2 under a novel asymptotic framework, where the number of folds K diverges to infinity as n diverges to infinity. Under this asymptotic framework, I show that DML2 offers theoretical advantages over DML1 in terms of bias and mean squared error (MSE). This result suggests that practitioners should adopt DML2 to achieve more accurate and reliable results. Additionally, it provides practical recommendations for improving DML2 performance in applications, specifically conditions under which setting K equals n minimizes asymptotic bias and MSE for DML2.

DML is useful in estimating a parameter θ_0 that 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), and treatment effect coefficient in the partial linear model (PLM), all of which have been studied in the literature on semi-parametric 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 a real-valued parameter θ_0, and the nuisance function η_0 consists of conditional expectations, such as the propensity score. This paper considers a setup that includes all these examples.

DML relies on two ingredients to guarantee that the estimation of θ_0 is as accurate as if the true η_0 had been used. The first ingredient is the *Neyman orthogonality* condition on the moment function m. This condition reduces the sensitivity of the estimation of θ_0 to errors in the estimation of η_0; see Remarks 3.1 and 3.2 for additional details. The second ingredient is the *cross-fitting* procedure, a sample-splitting method used to construct estimators for η_0. This procedure and the Neyman orthogonality condition on m remove the "own observation" bias, which arises when the same data are used to estimate both η_0 and θ_0.

DML1 and DML2 estimate θ_0 by first randomly dividing the data into K equal-sized folds, denoted by $\mathcal{I}_k$ 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, $K^{-1}\sum_{k=1}^{K}\left((n/K)^{-1}\sum_{i\in\mathcal{I}_k} m(W_i, \theta, \hat{\eta}_k)\right) = 0$.

Chernozhukov et al. (2018) conjectured that DML2 performs better than DML1 in small samples based on the simulation results of their relative performance. However, the existing asymptotic theory is insufficient for validating this conjecture, as it predicts that both DML1 and DML2 lead to estimators with the same limiting distribution, assuming that the number of folds K remains fixed as the sample size n diverges to infinity.

This paper studies the properties of the estimators based on DML1 and DML2 under a new asymptotic framework, aiming to understand which version has theoretical advantages. I consider an asymptotic framework where the number of folds $K \to \infty$ as $n \to \infty$. This framework offers a better description of finite sample situations where the practitioner implementing DML desires to increase K to improve the precision of the estimators $\hat{\eta}_k$'s, which use a fraction $(K-1)/K$ of the data. The use of alternative asymptotic approximations incorporating features of the finite sample problem is not new in the literature. An incomplete list of similar and recent approaches in different econometric problems includes Cattaneo and Jansson (2018), Bugni and Canay (2021), and Cai (2022), among many other authors.

This paper makes three contributions. First, it shows that DML2 offers theoretical advantages over DML1 in terms of bias and MSE, thereby formalizing a previous conjecture based on simulation evidence. More concretely, the first-order asymptotic distribution of DML1 may exhibit an asymptotic bias, which is not the case for DML2. This asymptotic bias is proportional to a parameter Λ that only depends on the moment function m, θ_0, and η_0. When Λ is equal to zero, DML1 and DML2 exhibit similar first-order asymptotic

properties. However, as Λ deviates from zero, DML1 becomes increasingly sensitive to large K values regarding bias and MSE, while DML2 remains unaffected by the choice of K. For several treatment effect parameters, such as ATE, ATT-DID, ATT, and PLM, Λ is equal to zero, but for others, like LATE and w-ATE, it is typically nonzero. The distinction between DML1 and DML2 through Λ emerges under the proposed asymptotic framework, providing insights not captured by the existing asymptotic theory or simulation-based evidence.

Second, this paper provides conditions that guarantee the asymptotic validity of implementing DML2 with any number of folds. Specifically, under these conditions, the first-order asymptotic distribution of the estimators based on DML2 when $K \to \infty$ as $n \to \infty$ is equal to the existing first-order asymptotic theory where K is fixed. In particular, the number of folds K can be set equal to the sample size n for DML2, obtaining the *leave-one-out* estimator with the same asymptotic distribution. This result is particularly useful for practitioners. It shows that dividing the data into many folds for DML2 implementation is asymptotically valid, which is a common practice to improve the precision of the nuisance function estimates $\hat{\eta}_k$'s; see Remark 2.3 for additional discussion on increasing the number of folds. Furthermore, the conditions guaranteeing the asymptotic validity for DML2 allow for the study of higher-order properties for all the DML2 estimators, including the leave-one-out estimator.

Third, this paper provides conditions under which the leave-one-out estimator is asymptotically optimal in terms of bias and MSE among the estimators based on DML2. More concretely, under these conditions, the absolute value of the leading term in the higher-order asymptotic bias of the DML2 estimators decreases as K increases, with the minimum achieved at $K = n$. Therefore, the leave-one-out estimator is optimal in terms of bias among the estimators based on DML2. It also holds that the leave-one-out estimator is optimal with respect to the second-order asymptotic MSE whenever certain data-dependent conditions hold, making it the most efficient choice for practitioners.

Finally, the previous results offer several lessons for practitioners. First, DML2 is the recommended option for DML implementation, especially in small-sample situations when increasing the number of folds is desired to improve the precision of the estimators $\hat{\eta}_k$. Second, choosing the number of folds K equal to the sample size n is optimal for DML2 implementation to reduce the asymptotic bias, where the asymptotic bias refers to the leading term of the higher-order asymptotic bias of the DML2 estimator. Third, choosing $K = n$ is optimal for the asymptotic accuracy of DML2 to estimate θ_0 when certain data-dependent condition holds, where the asymptotic accuracy refers to the second-order asymptotic MSE. The previous two lessons reveal that the common recommendations of choosing 5, 10 or 20 folds for the cross-fitting procedure in DML (e.g., Ahrens et al. (2024a,b), Bach et al. (2022), and Bach et al. (2024)) are suboptimal in terms of the asymptotic bias and the asymptotic

accuracy. The next lesson concerns the relative loss a practitioner can face by choosing a K that is different from the optimal choice. Fourth, if the optimal choice for minimizing the asymptotic bias and second-order asymptotic MSE is $K = n$, then choosing $K = 10$ to implement DML2 guarantees that the maximum relative loss compared to the optimal choice in terms of the asymptotic bias and the asymptotic accuracy is around 10% and 5%, respectively.

The conditions provided in this paper include stronger assumptions about the estimators of η_0 compared to the existing first-order asymptotic theory of DML to address two challenging situations. The first concerns the asymptotic properties of the estimators of θ_0, as the proof strategy for fixed K cannot be directly adapted to the case where $K \to \infty$ as $n \to \infty$. The second challenge involves analyzing the higher-order properties of the DML2 estimator when $K \to \infty$ as $n \to \infty$, which requires additional structure on the estimators of η_0.

Related Literature

This paper contributes to the growing literature on DML, where different estimators have been proposed based on DML for addressing semi-parametric estimation problems without requiring strong conditions on the estimators of η_0 (e.g., without invoking a Donsker class assumption). An incomplete list in this literature includes 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). All these papers used DML2 with some exceptions, such as Chernozhukov et al. (2017), Ji et al. (2023), and Cheng et al. (2023), which used DML1.[1] Except for Kennedy et al. (2024) and Jin and Syrgkanis (2024), all these papers derived the first-order asymptotic theory for their estimators, assuming that K remains fixed as $n \to \infty$. Kennedy et al. (2024) and Jin and Syrgkanis (2024) used a structure-agnostic framework to show the optimality of estimators based on DML. In contrast, I study the properties of estimators based on DML1 and DML2 when $K \to \infty$ as $n \to \infty$, showing that DML2 offers theoretical advantages over DML1 and providing conditions under which the leave-one-out estimator, defined as DML2 using $K = n$, is optimal in terms of asymptotic bias and MSE. To the best of my knowledge, this literature does not provide theoretical results on selecting K, which is also addressed as part of my results.

This paper also contributes to the literature on double-robust estimators, which includes Robins et al. (1994), Robins and Rotnitzky (1995), Scharfstein et al. (1999), Farrell (2015),

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

Sant'Anna and Zhao (2020), Chang (2020), Callaway and Sant'Anna (2021), Rothe and Firpo (2019), and Singh and Sun (2024), among others. Except Rothe and Firpo (2019), all these papers studied first-order asymptotic theory for their estimators that remain consistent even if some components of η_0 are misspecified. Rothe and Firpo (2019) studies the higher-order properties of double-robust estimators in a missing-data setting, where η_0 is estimated by a leave-one-out approach. My results complement their findings. First, the DML versions of the double-robust estimators allow a flexible estimation of the components of η_0 (e.g., nonparametric methods). Second, this paper presents the higher-order properties of estimators based on DML2. Third, among these DML2 estimators, the leave-one-out estimator is optimal in terms of bias and MSE whenever certain conditions hold. Interestingly, the leave-one-out estimator in the presented paper is the estimator studied in Rothe and Firpo (2019).

More broadly, this paper contributes to the literature on semi-parametric 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 of the papers in this literature provide conditions to study the estimators based on a plug-in approach (i.e., the same data are used to estimate η_0 and θ_0). In contrast, I provide conditions for studying the (higher-order) properties of estimators based on DML2.

Structure of the rest of the paper

The remainder of the paper is organized as follows. Section 2 describes the setup, notation, and estimators based on DML. Section 3 presents the formal results: Section 3.2 states the first-order asymptotic properties of the DML1 and DML2 estimators when $K \to \infty$ as $n \to \infty$, and Section 3.3 finds the higher-order properties of the DML2 estimators when $K \to \infty$ as $n \to \infty$. Section 4 presents the lessons for practitioners based on the formal results obtained in Section 3. Section 5 presents the Monte-Carlo simulations for ATT-DID (Sant'Anna and Zhao (2020)) and LATE (Hong and Nekipelov (2010)), using estimators based on DML1 and DML2 to examine the relevance of my asymptotic analysis in finite samples. Finally, Section 6 presents concluding remarks. Appendix A presents additional examples and results. Appendix B collects the proof of the main results. For brevity, the proofs of the auxiliary results appering in Appendix C are placed in the Online Appendix.[2]

[2] https://www.amilcarvelez.com/JMP/DML/online_appendix.pdf

2 Setup and Notation

This section presents the setup for the parameter of interest and the estimators based on DML. It contains examples previously studied in the literature that illustrate the setup. It also states the results of the existing asymptotic theory.

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

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

where $m : \mathcal{W} \times \Theta \times \mathcal{T} \to \mathbf{R}$ is a known moment function, $W \in \mathcal{W} \subseteq \mathbf{R}^{d_w}$ is a random vector with distribution F_0, and $X \in \mathcal{X} \subseteq \mathbf{R}^{d_x}$ is a sub-vector of W. The nuisance parameter $\eta_0 : \mathcal{X} \to \mathcal{T} \subseteq \mathbf{R}^p$ 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 $E[\psi^a(W, \eta_0(X))] \neq 0$ 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 partial derivative of the function 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 2.1), the average treatment effect on the treated in difference-in-differences designs (Example 2.2), and the local average treatment effect (Example 2.3), among others. Further examples are presented in Section A.1 of Appendix A.

Example 2.1 (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

$$\begin{aligned}
\psi^b(W, \eta) &= \eta_1 - \eta_2 + A(Y - \eta_1)\eta_3 - (1 - A)(Y - \eta_2)\eta_4 \ , \\
\psi^a(W, \eta) &= 1 \ ,
\end{aligned}$$

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 2.2 (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:

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

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

Example 2.3 (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:

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

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). □

2.1 Estimators based on DML

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

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

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

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

where $\eta_i = \eta_0(X_i)$ is the value of the nuisance parameter η_0 for the observation i, and X_i is a sub-vector of W_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.

For instance, an estimator $\hat{\eta}$ of η_0 can be obtained by using all the data, and then an esti-

mator 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 semi-parametric 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)); see also Remark 3.5. To attenuate the first-order effect of this bias on the plug-in estimators, stronger conditions are required on the estimator $\hat{\eta}$ (e.g., Donsker conditions). In contrast, DML —the approach considered in this paper— relies on general and simple conditions, such as a certain mean-square consistency condition, to obtain the standard properties (Chernozhukov et al. (2018), Chernozhukov et al. (2022a)).

In what follows, I explain how DML estimates η_0 and θ_0 by relying on cross fitting to avoids the "own observation" bias.

Estimates of the Nuisance Parameter

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:

1. *Sample splitting*: Randomly split the indices into K equal-sized folds $\mathcal{I}_k$, i.e., $\cup_{k=1}^{K} \mathcal{I}_k = \{1, 2, \ldots, n\}$. The number of observations in fold $\mathcal{I}_k$ is denoted $n_k = n/K$.[3]

2. *Nuisance Parameter 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 parameter $\eta_0(\cdot)$ 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 $k = 1, \ldots, K$.

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

[3]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.

DML1

The estimator based on DML1 first calculates preliminary estimators $\tilde{\theta}_k$ by solving the moment condition (1.1) within each fold $\mathcal{I}_k$ using the estimates $\hat{\eta}_i$,

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

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

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

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

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

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.

DML2

In contrast, the estimator based on DML2 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,

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

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

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

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

Remark 2.1. The estimators based on DML1 and DML2 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_0^a as in Example 2.1) 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,

$$\hat{\theta}_{n,1} = K^{-1} \sum_{k=1}^{K} \frac{n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^b(W_i, \hat{\eta}_i)}{\psi_0^a} = \frac{n^{-1} \sum_{i=1}^{n} \psi^b(W_i, \hat{\eta}_i)}{\psi_0^a} = \hat{\theta}_{n,2} \ .$$

Therefore, the DML1 and DML2 estimators for the ATE (Example 2.1) 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 $\hat{\theta}_{n,1} \neq \hat{\theta}_{n,2}$ in general. This occurs in all the other examples. □

Remark 2.2 (Oracle version of DML)**.** The oracle version of the DML1 estimator depends on random splitting, while this is not the case for the oracle version of the DML2. By oracle version, I refer to the case where the DML estimators are calculated assuming perfect knowledge of the η_i. More concretely, the oracle version of the estimator based on DML1 is defined as

$$\hat{\theta}^*_{n,1} = K^{-1} \sum_{k=1}^{K} \frac{n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^b(W_i, \eta_i)}{n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi^a(W_i, \eta_i)} \ , \tag{2.8}$$

which depends on sample splitting, i.e., the K-fold partition $\{\mathcal{I}_k : 1 \leq k \leq K\}$. In contrast, the oracle version of the estimator based on DML2 is defined as

$$\hat{\theta}^*_{n,2} = \frac{n^{-1} \sum_{i=1}^{n} \psi^b(W_i, \eta_i)}{n^{-1} \sum_{i=1}^{n} \psi^a(W_i, \eta_i)} \ , \tag{2.9}$$

which does not depend on sample splitting. Moreover, the oracle version of the DML2 estimator $\hat{\theta}^*_{n,2}$ is exactly the same as the one defined in (2.4), but this is not typically the case for the oracle version of the DML1 estimator $\hat{\theta}^*_{n,1}$ with some exceptions. When $\psi^a(W_i, \eta_i)$ has zero variance, then both $\hat{\theta}^*_{n,1}$ and $\hat{\theta}^*_{n,2}$ are equal to the one defined in (2.4). □

Remark 2.3. Simulation evidence has reported that increasing the number of folds K improves the performance of the estimators based on DML2 in terms of bias and mean squared error (Ahrens et al. (2024a,b) and Chernozhukov et al. (2018)). The cross-fitting procedure produces K possible different estimators $\hat{\eta}_k(\cdot)$ for the nuisance parameter $\eta_0(\cdot)$ and each of them uses a fraction $(K-1)/K$ of the data. For instance, these estimators use 50%, 80%, and 90% of the data when K is 2, 5, and 10, respectively. Therefore, the accuracy of these estimators increases with the values of K. 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. □

Previous Results

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

$$\sqrt{n}\left(\hat{\theta}_{n,j} - \theta_0\right) \xrightarrow{d} N(0, \sigma^2) \ , \tag{2.10}$$

where the variance of the asymptotic distribution is given by

$$\sigma^2 = \frac{E[m(W, \theta_0, \eta_0(X))^2]}{E[\psi^a(W, \eta_0(X))]^2} \ , \tag{2.11}$$

which only depends on the moment function m, the true nuisance parameter 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 for implementing DML.

The proof of (2.10) relies on a first-order equivalent condition. More concretely, both DML estimators $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ are first-order equivalent to the oracle estimator $\hat{\theta}_n^*$, which means

$$\sqrt{n}\left(\hat{\theta}_{n,j} - \hat{\theta}_n^*\right) \xrightarrow{p} 0 \ , \quad j = 1, 2 \ . \tag{2.12}$$

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

The first-order equivalence condition (2.12) was obtained when K is fixed as $n \to \infty$ by using (i) a Neyman orthogonality condition (which is necessary condition to obtain (2.12); see Remark 3.1) and (ii) a conditional independence property due to construction of the estimates $\hat{\eta}_i$ using the cross-fitting procedure (e.g., conditional on X_i, the estimation error $\hat{\eta}_i - \eta_i$ and W_i are independent). Importantly, the proof technique presented in Chernozhukov et al. (2018) relies on K being fixed as $n \to \infty$.

Although the existing asymptotic theory shows that DML1 and DML2 are asymptotically equivalent, DML2 is conjectured to perform better than DML1 based on the simulation results of their relative performance. To investigate whether DML2 offers theoretical advantages, this paper considers an asymptotic framework where $K \to \infty$ as $n \to \infty$ in the next section. There, it will be shown that the accuracy (bias and MSE) of DML1 is sensitive to K, while this is not the case for DML2, implying that DML2 asymptotically dominates DML1 in terms of bias and MSE. Furthermore, it presents conditions under which setting $K = n$ minimizes bias and MSE for the DML2 estimators, suggesting that practitioners should implement DML2 with $K = n$ in settings well-approximated by these conditions.

3 Main Results

This section presents the asymptotic properties of estimators based on DML1 and DML2 when $K \to \infty$ as $n \to \infty$.

3.1 Assumptions

This section presents and discusses the conditions required for the moment function and the nuisance function estimators. Assumption 3.1 specifies the formal conditions on the moment function m defined in (2.2), including a (strong) Neyman orthogonality condition, while Assumption 3.2 provides the details of the stochastic expansion satisfied by the nuisance parameter estimators. The technical conditions in parts (c) and (d) of Assumptions 3.2 are stronger than in the existing first-order asymptotic theory of DML to address the technical difficulties that arise when $K \to \infty$ as $n \to \infty$. Assumption 3.3 presents joint conditions on the moment function and the nuisance parameter estimators to conduct appropriate analysis of the leading terms of the higher-order bias and variance of the DML2 estimators.

The next assumption imposes conditions on the known functions ψ^a and ψ^b, which define the moment function m. These conditions are presented below and depend on the following finite positive constants M, C_0, C_1, C_2, and C_3.

Assumption 3.1. *The functions ψ^a and ψ^b are three-times continuously differentiable on $\eta \in \mathcal{T} \subseteq \mathbf{R}^p$ and satisfy for $z = a, b$,*

(a) $|E\left[\psi^a(W_i, \eta_i)\right]| > C_0$.

(b) $E\left[\partial_\eta \psi^z(W_i, \eta_i) \mid X_i\right] = 0$, *a.e.*

(c) $E\left[\psi^z(W_i, \eta_i)^4\right] < M$ *and* $E\left[||\partial_\eta \psi^z(W_i, \eta_i)||^4\right] < M$.

(d) $||E\left[(\partial_\eta \psi^z(W_i, \eta_i))(\partial_\eta \psi^z(W_i, \eta_i))^\top \mid X_i\right]||_\infty \leq C_1$.

(e) $\sup_{\eta \in \mathcal{T}} ||\partial_\eta^2 \psi^z(W_i, \eta)||_\infty \leq C_2$ *and* $\sup_{\eta \in \mathcal{T}} ||\partial_\eta^3 \psi^z(W_i, \eta)||_\infty \leq C_3$ *for* $z = a, b$.

where $\partial_\eta \psi^z(W_i, \eta_i)$ is the partial derivative of ψ^z with respect to η evaluated at $\eta_i = \eta_0(X_i)$ and the $||\cdot||_\infty$-norm is the maximum of the absolute value of the matrix entries.

Part (a) of Assumption 3.1 is an identification condition for the parameter θ_0. It implies that θ_0 can be written as a ratio of expected values as in (2.3). Part (b) of Assumption 3.1 guarantees that a (strong) *Neyman orthogonality* condition holds for the moment function m defined in (2.2). That is,

$$E\left[\partial_\eta m(W_i, \theta_0, \eta_i) \mid X_i\right] = 0 \ , \quad a.e. \tag{3.1}$$

The Neyman orthogonality condition is a necessary condition to guarantee first-order equivalence conditions, such as the one presented in (2.12); see Remark 3.1 for a further explanation. It has been used to remove the effects of the estimation of the nuisance parameter η_0 on the asymptotic distribution of estimators for θ_0. Many estimators of the treatment parameters of interest have associated moment functions that satisfy this condition, including the ones in Examples 2.1 (ATE), 2.2 (ATT-DID), and 2.3 (LATE). Weaker forms of this condition have been used in the literature for a similar purpose; see, for instance, Assumption 5.1 in Belloni et al. (2017), Assumption 3.1 in Chernozhukov et al. (2018), and Equation (2.12) in Andrews (1994). The Neyman orthogonality condition is helpful in studying the asymptotic properties of DML; however, it is not a restrictive requirement. Under certain conditions, it is possible to transform a moment function into a moment function that satisfies a Neyman orthogonality condition; see Remark 3.2 for additional details.

Part (c) of Assumption 3.1 is a regularity condition. It is used (i) to obtain a first-order equivalence property of the DML estimators and their oracle versions in Section 3.2 and (ii) to guarantee that the high-order asymptotic approximation and quantities that appear in Section 3.3 are well-defined. Parts (d) and (e) of Assumption 3.1 are mild technical conditions. It is possible to set $C_3 = 0$ when the moment function is a quadratic polynomial in η (the values of the nuisance parameter). This occurs when a doubly-robust moment condition defines the parameter of interest; see Theorem 4 in Chernozhukov et al. (2022a). All the examples presented in Section 2 and in Appendix A.1 satisfy Assumption 3.1.

Remark 3.1. Without suitable assumptions about the moment functions, a first-order equivalent condition between a feasible estimator and its oracle version (as in (2.12)) is not true in general. 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 $n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, \Sigma)$ and Σ is an invertible matrix. It can be shown that

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

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 (3.1) holds. □

Remark 3.2. Moment functions satisfying a Neyman orthogonality condition can be obtained by adding an adjustment term to the original moment functions. Specifically, under certain conditions on the nuisance parameter η_0, there exists α_0 (function of covariates X)

and $\phi(W, \theta, \eta, \alpha)$ (adjustment term) such that

- $E[\phi(W, \theta, \eta_0, \alpha_0)] = 0$
- the augmented moment function $\tilde{m}(W, \theta, \eta, \alpha) = m(W, \theta, \eta) + \phi(W, \theta, \eta, \alpha)$ satisfies a Neyman orthogonality condition

$$E\left[\partial_{\tilde{\eta}} \tilde{m}(W, \theta_0, \tilde{\eta})|_{\tilde{\eta}=\tilde{\eta}_0(X)} \mid X\right] = 0 \ , \quad a.e.$$

  where $\tilde{\eta}_0(X) = (\eta_0(X), \alpha_0(X))$.

The augmented term ϕ is an influence function of a particular parameter. It can be obtained using methods previously developed in the literature, such as Ichimura and Newey (2022) and Newey (1994). Recently, Chernozhukov et al. (2022a) used the approach of Ichimura and Newey (2022) and estimators based on DML2 to propose debiased GMM estimators. □

Stochastic Expansion for the Nuisance Parameter Estimator

The next assumption imposes additional structure on the nuisance parameter estimators compared to the existing DML framework. These stronger conditions have a twofold purpose: addressing the technical challenges that arise when $K \to \infty$ as $n \to \infty$ in Section 3.2 and allowing the analysis of higher-order properties in Section 3.3.

To fix ideas, consider an estimator $\hat{\eta}$ of η_0 at a given point x such that (i) $n^{-2\varphi_1}$ is the convergence rate of the variance of $\hat{\eta}$ and (ii) $n^{-\varphi_2}$ is the convergence rate of the bias of $\hat{\eta}$. To have more information about the variance and bias part, the next assumption imposes an asymptotic linear representation for both parts. More concretely, it takes as given the positive constants φ_1 and φ_2 and assumes the existence of two sequences of functions δ_n (for the variance part) and b_n (for the bias part) that satisfy the asymptotic linear representation. Additional assumptions on these functions are imposed and depend on the following finite positive constants M_1, C_δ, and C_b, and a sequence of positive constants τ_n converging to zero (i.e., $\tau_n = o(1)$). Before continuing, let $n_0 = ((K-1)/K)n$ be the number of observations in the sample $\{W_i : i \notin \mathcal{I}_k\}$ used by $\hat{\eta}_k(\cdot)$ to estimate $\eta_0(\cdot)$.

Assumption 3.2. *There exist two sequences of functions $\delta_{n_0} : \mathcal{W} \times \mathcal{X} \to \mathbf{R}^p$ and $b_{n_0} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}^p$, such that*

(a) For any given $x \in \mathcal{X}$ and $k \in \{1, \ldots, K\}$,

$$\hat{\eta}_k(x) - \eta_0(x) = n_0^{-1/2} \sum_{\ell \notin \mathcal{I}_k} n_0^{-\varphi_1} \delta_{n_0}(W_\ell, x) + n_0^{-1} \sum_{\ell \notin \mathcal{I}_k} n_0^{-\varphi_2} b_{n_0}(X_\ell, x) + n_0^{-2\min\{\varphi_1, \varphi_2\}} \hat{R}(x) \ ,$$

where $\hat{R}(x) = O_p(1)$, $E[\delta_{n_0}(W_\ell, x) \mid X_\ell] = 0$ *a.e., and* $E\left[||\delta_{n_0}(W_\ell, x)||^2\right] > C_\delta$.

(b) For any $i \neq j$,

(b.1) $E\left[E[||\delta_{n_0}(W_j, X_i)||^2 \mid X_i]^2\right] \leq M_1$ *and* $E\left[E\left[||n_0^{-\varphi_2} b_{n_0}(X_j, X_i)||^2 \mid X_i\right]^2\right] \leq n_0^{2(1-2\varphi_1)} \tau_{n_0}$.

(b.2) $E\left[||\delta_{n_0}(W_j, X_i)||^{2s}\right] < n_0^{(s-1)(1-2\varphi_1)} M_1$ *for* $s = 1, 2$.

(b.3) $E\left[||E\left[b_{n_0}(X_j, X_i) \mid X_i\right]||^4\right] \in (C_b, M_1)$.

(b.4) $E\left[||n_0^{-\varphi_2} b_{n_0}(X_j, X_i)||^{2s}\right] < n_0^{(2s-1)(1-2\varphi_1)} \tau_{n_0}$ *for* $s = 1, 2$.

(c) $E[||\hat{R}(X_i)||^2] = O(1)$.

(d) $n^{-1} \sum_{i=1}^n ||\hat{R}(X_i)||^4 = O_p(n^{4\min\{\varphi_1, \varphi_2\}})$.

Part (a) of Assumption 3.2 present a stochastic expansion for the estimation error of the nuisance parameter estimator $\hat{\eta}_k(x)$. It assumes that this approximation has two terms that model the variance ($n_0^{-1/2} \sum_{\ell \notin \mathcal{I}_k} n_0^{-\varphi_1} \delta_{n_0}(W_\ell, x)$) and bias ($n_0^{-1} \sum_{\ell \notin \mathcal{I}_k} n_0^{-\varphi_2} b_{n_0}(X_\ell, x)$) components of the estimator $\hat{\eta}_k(x)$. Nonparametric kernel estimators satisfy this asymptotic expansion under mild regularity conditions on the nuisance parameter η_0. Appendix A.3 presents δ_{n_0} and b_{n_0} for a class of nonparametric kernel estimators and the nuisance parameters η_0 (conditional expectations) that appear in the examples.

Part (b) of Assumption 3.2 presents regularity conditions on $\delta_{n_0}(W_j, X_i)$ and $b_{n_0}(X_j, X_i)$ for $j \neq i$. Part (b.1) is helpful to establish that the leading terms in the stochastic approximation for the estimation error of $\hat{\eta}_k$ have a finite fourth moment; see Lemma C.1 in Appendix C. Part (b.2) and the condition $E\left[||\delta_{n_0}(W_\ell, x)||^2\right] > C_\delta$ in part (a) guarantee that the variance of the nuisance parameter estimator has a convergence rate $O(n^{-2\varphi_1})$, while part (b.3) establishes that the bias term has a convergence rate $O(n^{-\varphi_2})$. Finally, part (b.4) considers additional regularity conditions. These regularity conditions can be verified for Nadaraya-Watson estimators and the nuisance parameters η_0 (conditional expectations) that appear in the examples.

Parts (c) and (d) of Assumption 3.2 are helpful high-level conditions to establish results in Section 3.2, where the asymptotic framework considers $K \to \infty$ as $n \to \infty$. This assumption can be verified for Nadaraya-Watson estimators under additional suitable conditions on the nuisance parameter η_0 (conditional expectations) that appear in the examples. Part (c) is sufficient to guarantee that $E[||\hat{\eta}_i - \eta_i||^2]$ is finite and has convergence rate $O(n^{-2\varphi_1}) + O(n^{-2\varphi_2})$. Part (d) is sufficient to guarantee that $n^{-1} \sum_{i=1}^n ||\hat{\eta}_i - \eta_i||^4$ is $O_p(n^{-4\min\{\varphi_1, \varphi_2\}})$. Both intermediate results are formally established in Lemma C.4 in Appendix C.

Stochastic Expansion for the DML2 Estimator

Finally, the next assumption imposes joint conditions on the functions δ_{n_0} and b_{n_0}, defined in Assumption 3.2, and the moment functions m, defined in (2.2). These conditions are important to (i) derive a valid stochastic expansion for the estimators based on DML2 and (ii) conduct an appropriate analysis of the leading terms of the higher-order bias and variance of the DML2 estimator. Before continuing, define $\tilde{b}_{n_0}(X_i) = E\left[b_{n_0}(X_j, X_i) \mid X_i\right]$ for $j \neq i$.

Assumption 3.3.

(a) The following limits exist and are finite,

$$G_\delta = \lim_{n_0\to\infty} E\left[E\left[\delta_{n_0}(W_j, X_i)^\top \left(\partial_\eta^2 m(W_i, \theta_0, \eta_i)\right) \delta_{n_0}(W_\ell, X_i) \mid W_j, W_\ell\right]^2\right] / E[\psi^a(W_i, \eta_i)]^2, \tag{3.2}$$

$$F_\delta = \lim_{n_0\to\infty} \frac{1}{2} E\left[\delta_{n_0}(W_j, X_i)^\top \left(\partial_\eta^2 m(W_i, \theta_0, \eta_i)\right) \delta_{n_0}(W_j, X_i)\right] / E[\psi^a(W_i, \eta_i)] \tag{3.3}$$

$$F_b = \lim_{n_0\to\infty} \frac{1}{2} E\left[\tilde{b}_{n_0}(X_i)^\top \left(\partial_\eta^2 m(W_i, \theta_0, \eta_i)\right) \tilde{b}_{n_0}(X_i)\right] / E[\psi^a(W_i, \eta_i)] \ , \tag{3.4}$$

$$G_b = \lim_{n_0\to\infty} E\left[m(W_j, \theta_0, \eta_j)\delta_{n_0}(W_j, X_i)^\top \left(\partial_\eta^2 m(W_i, \theta_0, \eta_i)\right) \tilde{b}_{n_0}(X_i)\right] / E[\psi^a(W_i, \eta_i)]^2 \ , \tag{3.5}$$

(b) $G_\delta > 0$, $F_\delta + F_b \neq 0$, and $G_b \neq 0$.

Part (a) of Assumption 3.3 is a regularity condition. Assumptions 3.1 and 3.2 ensure that the sequences on the right-hand side of (3.2)–(3.5) are bounded, implying that, if the limit exists, it is finite. Whether the limits in (3.2)–(3.5) exist depends on the estimator of the nuisance function, η_0. For instance, this can be verified for Nadaraya-Watson estimators —using the same kernel as $n_0 \to \infty$— under additional suitable conditions on η_0. The limit may not exist when η_0 is estimated by two different types of estimators based on n_0, which is the number of observations in the estimation of η_0. For instance, if the estimator of η_0 uses a given kernel when n_0 is odd and a different one when n_0 is even, the limit can exist for each subsequence, but they may be different.

Part (b) of Assumption 3.3 is a sufficient condition to conduct an appropriate analysis of the leading terms of the higher-order bias and variance of the DML2 estimators in Section 3.3. Determining whether part (b) of Assumption 3.3 hold may require a case-by-case analysis, as it depends on the interaction between the second-order partial derivatives of the moment function m with respect to η and the estimator of the nuisance function. Importantly, a necessary condition is that the moment function m is a nonlinear function on η, i.e., the

matrix of second-order partial derivatives of m with respect to η is different than zero. For Examples 2.1 (ATE), 2.2 (ATT-DID), A.2 (ATT), and A.3 (PLM), it can be verified that $G_\delta > 0$, $|F_\delta + F_b| > 0$, and $G_b \neq 0$ when η_0 is estimated using Nadaraya-Watson estimators.

3.2 A First-Order Asymptotic Theory when K increases

This section presents the asymptotic distribution of the estimators based on DML1 and DML2 in an asymptotic framework where the number of folds K increases with the sample size n. Specifically, it shows that DML1 may exhibit a first-order asymptotic bias (i.e., its asymptotic distribution may not be centered at zero), which is not the case for DML2. These results show that DML2 offers theoretical advantages over DML1 in terms of bias and MSE. Furthermore, the conditions used in this section guarantee that DML2 remains asymptotically valid for any $K \leq n$.

The asymptotic framework considered in this section for DML is, to the best of my knowledge, new and approximates finite sample situations often faced by practitioners. For instance, small-sample situations where the practitioner implementing DML desires to increase K to improve the precision of the nuisance parameter estimators $\hat{\eta}_k$'s, which use a fraction $(K-1)/K$ of the data; see Remark 2.3. This finite sample situation is not well approximated by the available asymptotic framework in the literature, which considers that K is fixed as $n \to \infty$.

The next results present the asymptotic distributions of the estimators based on DML1 and DML2 when the number of folds $K \to \infty$ as $n \to \infty$.

Theorem 3.1. *Suppose Assumptions 3.1 and 3.2 hold. In addition, assume K is such that $K \leq n$, $K \to \infty$ and $K/\sqrt{n} \to c \in [0, \infty)$ as $n \to \infty$. If $\varphi_1 \leq 1/2$ and $1/4 < \min\{\varphi_1, \varphi_2\}$, then*

$$n^{1/2}\left(\hat{\theta}_{n,1} - \theta_0\right) \xrightarrow{d} N(c\Lambda, \sigma^2) \ ,$$

where $\hat{\theta}_{n,1}$ and σ^2 are as in (2.6), (2.11), respectively, and

$$\Lambda = \frac{Cov\left[m(W_i, \theta_0, \eta_i), \ -\psi^a(W_i, \eta_i)\right]}{E[\psi^a(W_i, \eta_i)]^2} \ . \tag{3.6}$$

Theorem 3.2. *Suppose Assumptions 3.1 and 3.2 hold. In addition, assume K is such that $K \leq n$, $K \to \infty$ as $n \to \infty$. If $\varphi_1 \leq 1/2$ and $1/4 < \min\{\varphi_1, \varphi_2\}$, then*

$$n^{1/2}\left(\hat{\theta}_{n,2} - \theta_0\right) \xrightarrow{d} N(0, \sigma^2) \ ,$$

where $\hat{\theta}_{n,2}$ and σ^2 are as in (2.7) and (2.11), respectively.

Theorems 3.1 and 3.2 explain why DML1 and DML2 behave similarly in simulations for many models previously studied in the literature, including the ones presented in Example 2.1 (ATE), Example 2.2 (ATT-DID), and Example A.3 (PLM). All these examples have $\Lambda = 0$, and, therefore, by these theorems, it follows that both estimators have the same asymptotic distribution whenever $K \to \infty$ slowly as $n \to \infty$ (i.e., $K = O(n^{1/2})$).

Theorem 3.1 shows that (first-order) asymptotic properties of DML1 can be sensitive to K when $\Lambda \neq 0$. Intuitively, this theorem shows the distribution of $n^{1/2}\left(\hat{\theta}_{n,1} - \theta_0\right)$ can be approximated by $N(\Lambda K/\sqrt{n}, \sigma^2)$, which is sensitive to the choice of K when n is small. In particular, when $K \sim \sqrt{n}$ and $\Lambda \neq 0$, the asymptotic distribution of estimators based on DML1 is not centered at zero, i.e., there is an asymptotic bias proportional to Λ affecting the reliability of the inference procedure and the accuracy of the estimation. Some examples where Λ is typically nonzero are Example 2.3 (LATE) and Example A.1 (w-ATE).

The previous implications make clear that Λ is a *discrepancy measure* between DML1 and DML2. More concretely, when Λ equals zero, DML1 and DML2 exhibit similar first-order asymptotic properties, but as Λ deviates from zero, DML1 becomes more sensitive to large values of K in terms of bias and MSE, while this is not the case of DML2. See Remark 3.6 for additional discussion on why DML1 has asymptotic bias proportional to Λ.

The conditions presented in Theorem 3.2 guarantee that estimators based on DML2 are asymptotically valid for any $K \leq n$. More concretely, under the conditions presented in this theorem, all the DML2 estimators using different K share the same asymptotic distribution $N(0, \sigma^2)$. This class of DML2 estimators includes the *leave-one-out* estimator defined by setting $K = n$. These results have practical consequences for the practitioner that desires to increase K to improve the precision of the nuisance parameter estimators. This theorem indicates that these DML2 estimators obtained by increasing K are asymptotically valid.

Finally, the proofs of Theorems 3.1 and 3.2 share the same structure and rely on two intermediate results. The first intermediate result is presented in Theorem C.1 for DML1 and in Theorem C.2 for DML2. These theorems state that a first-order equivalence condition holds for the estimators based on DML and their oracle versions defined in Remark 2.2 when $K \to \infty$ as $n \to \infty$. The second intermediate result calculates the asymptotic distribution of the oracle version of the DML estimators defined in Remark 2.2. This result is presented in Proposition C.1. It can be shown that Λ appears as an asymptotic bias for DML1 because the oracle version of the DML1 estimators depends on sample splitting, while this is not the case for the oracle version of the DML2 estimators.

Remark 3.3. If $K/\sqrt{n} \to \infty$ and $\Lambda > 0$, then the estimators based on DML1 may have a degenerate asymptotic distribution. More concretely, when $K \sim n^{1/2+\epsilon}$ for some sufficiently small $\epsilon > 0$, it is possible to guarantee that (i) $n^{1/2}\left(\hat{\theta}_{n,1} - \hat{\theta}^*_{n,1}\right) = o_p(1)$ by extending

Theorem C.1 and Lemma C.3, and (ii) $n^{1/2}\left(\hat{\theta}^*_{n,1} - \theta_0\right)$ converge to ∞ with probability approaching one, which follows by the proof of Proposition C.1. Therefore, $n^{1/2}\left(\hat{\theta}_{n,1} - \theta_0\right)$ can converge to ∞, which is a degenerate distribution. □

Remark 3.4. The first-order equivalence condition between $\hat{\theta}_{n,j}$ and its oracle version $\hat{\theta}^*_{n,j}$ relies on stronger assumptions compared to the existing DML framework to accommodate that $K \to \infty$ and $n \to \infty$. These assumptions are presented in parts (c) and (d) of Assumption 3.2, and they are used to prove two important intermediate results that appear in Lemma C.2,

$$n^{-1/2} \sum_{i=1}^{n} (\hat{\eta}_i - \eta_i)^\top \partial_\eta m(W_i, \theta_0, \eta_i) = o_p(1) \ , \tag{3.7}$$

and in Lemma C.3,

$$\max_{k=1,\ldots,K} \left| n_k^{-1/2} \sum_{i \in \mathcal{I}_K} (\hat{\eta}_i - \eta_i)^\top \partial_\eta m(W_i, \theta_0, \eta_i) \right| = o_p(1) \ . \tag{3.8}$$

When K is fixed as $n \to \infty$, the previous intermediate results, (3.7) and (3.8), follow from a Bonferroni correction argument and

$$\left| n_k^{-1/2} \sum_{i \in \mathcal{I}_K} (\hat{\eta}_i - \eta_i)^\top \partial_\eta m(W_i, \theta_0, \eta_i) \right| = o_p(1) \ . \tag{3.9}$$

The proof of (3.9) when K is fixed follows by (i) a Neyman orthogonality condition and (ii) a conditional independence property due to the construction of the estimates $\hat{\eta}_i$ using cross-fitting (e.g., conditional on X_i, the estimation error $\hat{\eta}_i - \eta_i$ and W_i are independent). However, this proof cannot be adapted to the case where $K \to \infty$ as $n \to \infty$. □

Remark 3.5. The condition (3.7) is important to obtain asymptotically valid estimators with the same asymptotic distribution as the oracle estimator in (2.4). Remark 3.4 pointed out this is the case for DML estimators. Equivalent formulations of (3.7) as a high-level condition have been used in the literature to establish a first-order equivalence condition between a plug-in estimator—defined in Section 2.1— and its oracle version (e.g., Andrews (1994), Farrell (2015)). In general, the verification of (3.7) for plug-in estimators is difficult since it is unclear whether $E[(\hat{\eta}_i - \eta_i)^\top \partial_\eta m(W_i, \theta_0, \eta_i)]$ is zero (or asymptotically zero) due to the correlation between $\hat{\eta}_i - \eta_i$ and $\partial_\eta m(W_i, \theta_0, \eta_i)$, which is a manifestation of the "own observation" bias that arise because the same data are used to estimate η_0 and θ_0. □

Remark 3.6. The discrepancy measure Λ is proportional to the first-order asymptotic bias of the DML1 estimator because its oracle version depends on sample splitting, which is

not the case for the DML2 estimator. For the sake of explanation, suppose the nuisance parameter is known. In this case, the oracle estimator $\hat{\theta}^*_{n,1}$ defined in Remark 2.2 is equal to the average of K preliminary estimators $\tilde{\theta}^*_k$, that is

$$\hat{\theta}^*_{n,1} = K^{-1} \sum_{k=1}^{K} \tilde{\theta}^*_k \ ,$$

where each $\tilde{\theta}^*_k$ is as in (2.4) but using only observations in the fold $\mathcal{I}_k$, which has n/K observations. Therefore, each of these preliminary estimators has a (higher-order) asymptotic bias equal to $\Lambda(n/K)^{-1}$ since it uses n/K observations. An explicit expression for Λ can be obtained based on standard arguments (e.g., Newey and Smith (2004)). Since the bias of the average of the estimators is the same as the average of the bias of the estimators, it follows that the (higher-order) asymptotic bias of $\hat{\theta}^*_{n,1}$ is $\Lambda K/n$. Intuitively, when $K \sim \sqrt{n}$, this asymptotic bias become proportional to $\Lambda/\sqrt{n}$ and shows up in the first-order asymptotic distribution of the oracle estimator $\hat{\theta}^*_{n,1}$. Finally, since the feasible DML1 estimator $\hat{\theta}_{n,1}$ and $\hat{\theta}^*_{n,1}$ are first-order equivalent, their asymptotic distributions are the same. □

3.3 High-Order Asymptotic Theory for DML2 estimators

This section presents higher-order asymptotic properties (e.g., bias, variance, and MSE) of the estimators based on DML2 in an asymptotic framework where the number of folds K increases with the sample size n. Specifically, it shows that the leading term of the higher-order bias is a decreasing function of K. Moreover, it presents conditions under which setting K equals n minimizes the second-order MSE of DML2 estimators.

The goal of this section is to propose asymptotic approximations that offer a better description of the finite sample behavior of the DML2 estimators. The approximations based on the first-order asymptotic distribution are insufficient for this goal since all the DML2 estimators share the same asymptotic distribution for any $K \leq n$ (Theorem 3.2). In this section, I obtain better approximations by considering *stochastic expansions* up to a smaller remainder error term than in the existing first-order asymptotic theory.

The main idea is to use these better approximations to study the higher-order asymptotic properties of the estimators based on DML2, with the hope that they are reliable enough to explain the finite sample behavior of the estimators. This approach has a long history in econometrics to compare estimators that are first-order equivalent (e.g., Rothenberg (1984), Linton (1995), Newey and Smith (2004), Graham et al. (2012)).

Stochastic Expansions for DML2 estimators

The next result presents a stochastic expansion for the estimators based on DML2 that is asymptotically valid when K increases with the sample size n. This stochastic expansion focuses on the nuisance parameter estimators satisfying Assumption 3.2 with $\varphi_1 = \varphi_2$; see Remark 3.8 for additional discussion of other cases. For the remainder of this section, I set $\varphi = \varphi_1 = \varphi_2$.

Theorem 3.3. *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi \in (1/4, 1/2)$, then*

$$n^{1/2}(\hat{\theta}_{n,2} - \theta_0) = \mathcal{T}_n^* + \mathcal{T}_{n,K}^{nl} + R_{n,K} \ , \tag{3.10}$$

where $\hat{\theta}_{n,2}$ is as (2.7), $\mathcal{T}_n^*$ is defined in (3.11) and $\mathcal{T}_n^* \xrightarrow{d} N(0, \sigma^2)$ as $n \to \infty$ with σ^2 defined as in (2.11), $T_{n,K}^{nl}$ is defined in (3.12) and satisfies (i) $\lim_{n\to\infty} \inf_{K\leq n} Var[n^{2\varphi-1/2} T_{n,K}^{nl}] > 0$ and (ii) $\lim_{n\to\infty} \sup_{K\leq n} E\left[\left(n^{2\varphi-1/2} T_{n,K}^{nl}\right)^2\right] < \infty$, and

$$\lim_{n\to\infty} \sup_{K\leq n} P(n^{2\varphi-1/2}|R_{n,K}| > \epsilon) = 0 \ ,$$

for any given $\epsilon > 0$.

Theorem 3.3 presents a stochastic expansion more accurate than the available first-order asymptotic theory for any given sequence $K \to \infty$ as $n \to \infty$. More concretely, equation (3.10) presents a remainder error term $R_{n,k}$ that is stochastically smaller than $T_{n,K}^{nl}$, i.e., $n^{2\varphi-1/2} R_{n,k}$ converges to zero in probability for any sequence $K \to \infty$, while this is not the case for $n^{2\varphi-1/2} T_{n,k}^{nl}$ since its variance is positive for any n sufficiently large. Therefore, under the conditions of this theorem, $R_{n,K}$ in (3.10) is smaller than the remainder error term $R_{n,K}^* = T_{n,K}^{nl} + R_{n,K}$ obtained by the first-order approximation, denoted by $\mathcal{T}_n^*$,

$$n^{1/2}(\hat{\theta}_{n,2} - \theta_0) = \mathcal{T}_n^* + R_{n,K}^* \ ,$$

where

$$\mathcal{T}_n^* = n^{-1/2} \sum_{n=1}^{n} m(W_i, \theta_0, \eta_i) / E[\psi^a(W_i, \eta_i)] \ . \tag{3.11}$$

The approximation in Theorem 3.3 includes the additional term $\mathcal{T}_{n,K}^{nl}$ to accommodate for the errors of the nuisance parameter estimators. More concretely, under the conditions of Theorem 3.3, $\mathcal{T}_{n,K}^{nl}$ is defined as the leading term in the scaled difference between the feasible

estimator $\hat{\theta}_{n,2}$ and oracle estimators $\hat{\theta}^*_{n,2}$ defined in Remark 2.2,

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

where

$$\lim_{n\to\infty} \sup_{K\leq n} P(n^{2\varphi-1/2}|R^{nl}_{n,K}| > \epsilon) = 0$$

and

$$\mathcal{T}^{nl}_{n,K} = \frac{1}{2} n^{-1/2} \sum_{k=1}^{K} \sum_{i\in\mathcal{I}_k} \Delta_i^\top \left(\partial^2_\eta m(W_i, \theta_0, \eta_i)/E[\psi^a(W_i, \eta_i)]\right) \Delta_i \ , \tag{3.12}$$

with Δ_i defined below for $i \in \mathcal{I}_k$,

$$\Delta_i = n_0^{-1/2} \sum_{\ell\notin\mathcal{I}_k} n_0^{-\varphi} \delta_{n_0}(W_\ell, X_i) + n_0^{-1} \sum_{\ell\notin\mathcal{I}_k} n_0^{-\varphi} b_{n_0}(X_\ell, X_i) \ , \tag{3.13}$$

where δ_{n_0} and b_{n_0} are the functions in Assumption 3.2. The approximation of the scaled difference between $\hat{\theta}_{n,2}$ and $\hat{\theta}^*_{n,2}$ is obtained by using Taylor expansions to approximate $\psi^z(W_i, \hat{\eta}_i)$ by $\psi^z(W_i, \eta_i)$ for $z = a, b$.

Remark 3.7. When K is fixed as $n \to \infty$, a similar stochastic expansion can be derived for DML1,

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

Furthermore, the stochastic expansion in (3.10) for DML2 remains valid when K is fixed as $n \to \infty$. These expressions show that when K is fixed and $\varphi = \varphi_1 = \varphi_2$, the two leading terms in the stochastic approximation are the same. □

Remark 3.8. Theorem 3.3 presents a stochastic expansion for the case $\varphi = \varphi_1 = \varphi_2$, representing situations where the nuisance function estimator balances bias and variance. For example, this occurs when a bandwidth with an optimal convergence rate is used in Nadaraya-Watson estimators to estimate the nuisance function. Appendix A.2 discusses the case where $\varphi_1 < \varphi_2$, which represents situations where the bias of the nuisance function estimator converges faster than the variance components (e.g., *undersmoothing*). Theorem A.1 extends Theorem 3.3 by providing the stochastic expansion for the alternative cases of (φ_1, φ_2). □

High-Order Asymptotic Properties

I calculate the higher-order asymptotic bias, variance, and mean squared error (MSE) of the estimator $\hat{\theta}_{n,2}$ by using the asymptotic approximation $\mathcal{T}_{n,K}$ defined next,

$$\mathcal{T}_{n,K} = \mathcal{T}_n^* + \mathcal{T}_{n,K}^{nl} \ . \tag{3.14}$$

More concretely, the higher-order bias, variance, and MSE of $\hat{\theta}_{n,2}$ are respectively defined as

$$\begin{aligned}
\text{HO-Bias}[\hat{\theta}_{n,2}] &= n^{-1/2} E[\mathcal{T}_{n,K}] \ , \\
\text{HO-Var}[\hat{\theta}_{n,2}] &= n^{-1} \text{Var}[\mathcal{T}_{n,K}] \ , \\
\text{HO-MSE}[\hat{\theta}_{n,2}] &= n^{-1} E[\mathcal{T}_{n,K}^2] \ .
\end{aligned}$$

Theorems 3.4 and 3.5, and Corollary 3.1 present explicit expressions for the leading terms of $E[\mathcal{T}_{n,K}]$, $\text{Var}[\mathcal{T}_{n,K}]$, and $E[\mathcal{T}_{n,K}^2]$.

Similar definitions of higher-order asymptotic bias and variance have been used to compare alternative estimators with the same asymptotic distribution, including Rothenberg (1984), Linton (1995), and Newey and Smith (2004). As discussed in Rothenberg (1984), these definitions are valid asymptotic approximations of the bias and variance of the estimators whenever additional regularity conditions hold. An alternative interpretation discussed in Linton (1995) suggests that these definitions could be interpreted as a form of approximations of the bias and variance of $\hat{\theta}_{n,2}$ since they are based on the moments of the approximation $\mathcal{T}_{n,K}$, which has a distribution that asymptotically approximates the distribution of $n^{1/2}(\hat{\theta}_{n,2} - \theta_0)$ up to an error $o(n^{1/2-2\varphi})$ under certain regularity conditions.

The next theorem presents the expected value and variance of $\mathcal{T}_{n,K}$, which can be used to calculate the higher-order bias and variance of the DML2 estimator.

Theorem 3.4 (Higher-Order Bias)**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi \in (1/4, 1/2)$, then*

$$E[\mathcal{T}_{n,K}] = (F_\delta + F_b) \left(1 + \frac{1}{K-1}\right)^{2\varphi} n^{1/2-2\varphi} + \nu_{n,K} \ ,$$

where

$$\sup_{K \leq n} |\nu_{n,K}| = o(n^{1/2-2\varphi}) \ ,$$

with $\mathcal{T}_{n,K}$, F_δ and F_b defined as in (3.14), (3.3) and (3.4), respectively.

Theorem 3.5 (Higher-Order Variance). *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi \in (1/4, 1/2)$, then*

$$Var[\mathcal{T}_{n,K}] = \sigma^2 + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi - 1/2} n^{1/2-2\varphi} + r_{n,K} \ ,$$

where

$$\sup_{K \leq n} |r_{n,K}| = o(n^{1/2-2\varphi}) \ ,$$

with $\mathcal{T}_{n,K}$, σ^2, and G_b defined as in (3.14), (2.11), and (3.5), respectively.

Theorems 3.4 and 3.5 can be used to find the higher-order bias and variance of $\hat{\theta}_{n,2}$,

$$\text{HO-Bias}[\hat{\theta}_{n,2}] = (F_\delta + F_b) \left(1 + \frac{1}{K-1}\right)^{2\varphi} n^{-2\varphi} + \nu_{n,K} n^{-1/2} \ ,$$

and

$$\text{HO-Var}[\hat{\theta}_{n,2}] = \sigma^2 n^{-1} + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi - 1/2} n^{-1/2-2\varphi} + r_{n,K} n^{-1} \ .$$

The leading term of the higher-order bias depends on F_δ and F_b defined in (3.3) and (3.4), respectively. While the second leading term of the higher-order variance depends on G_b defined in (3.5). These are quantities that depend on three elements: (i) the functions δ_{n_0} and b_{n_0} that appear in the stochastic expansion for the nuisance parameter estimator presented in Assumption 3.2, (ii) the second-order derivatives of the moment function m with respect to η, and (iii) the data distribution.

The previous calculation reveals that the absolute value of the leading term in the higher-order bias, $|F_\delta + F_b|(1 + 1/(K-1))^{2\varphi} n^{-2\varphi}$, decreases as K increases. Therefore, the leave-one-out estimator, defined as the DML2 estimator with $K = n$, minimizes the absolute value of the leading term in the higher-order asymptotic bias. The two leading terms of the higher-order variance, $\sigma^2 n^{-1}$ and $G_b(K/(K-1))^{2\varphi - 1/2} n^{-1/2-2\varphi}$, define a decreasing function of K when $G_b > 0$. Similarly, when $G_b > 0$, the leave-one-out minimize the two leading terms of the higher-order variance.

The next result is a corollary derived from Theorems 3.4 and 3.5. It presents the second moment of $\mathcal{T}_{n,K}$, which can be used to calculate the higher-order MSE of the DML2 estimator.

Corollary 3.1. *Suppose the conditions of Theorems 3.4 and 3.5 holds. Then,*

$$E[\mathcal{T}_{n,K}^2] = \sigma^2 + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi - 1/2} n^{1/2-2\varphi} + \tilde{r}_{n,K} \ ,$$

where

$$\sup_{K\leq n} |\tilde{r}_{n,K}| = o(n^{1/2-2\varphi}) \ ,$$

with $\mathcal{T}_{n,K}$, σ^2 *and* G_b *defined as in* (3.14), (2.11) *and* (3.5), *respectively.*

Corollary 3.1 can be used to find the higher-order MSE,

$$\text{HO-MSE}[\hat{\theta}_{n,2}] = \sigma^2 n^{-1} + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi-1/2} n^{-1/2-2\varphi} + \tilde{r}_{n,K} n^{-1} \ .$$

The two leading terms of HO-MSE$[\hat{\theta}_{n,2}]$ define *the second-order asymptotic MSE*,

$$\text{SO-MSE}[\hat{\theta}_{n,2}] = \sigma^2 n^{-1} + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi-1/2} n^{-1/2-2\varphi} \ ,$$

which is a decreasing function on K when G_b is positive. Therefore, the leave-one-out estimator can minimize the second-order asymptotic MSE by setting $K = n$ when $G_b > 0$.

Remark 3.9. When K is fixed as $n \to \infty$, it can be shown that the two leading terms of the higher-order MSE, σ^2/n and $G_b \left(1 + 1/(K-1)\right)^{2\varphi-1/2} n^{-1/2-2\varphi}$, are the same for DML1 and DML2 up to an error of size $o(n^{-1/2-2\varphi})$. For this conclusion is important that K is fixed as $n \to \infty$, since some higher-order terms in the asymptotic approximation for DML1 may become large terms for larger values of K. Remark 3.10 illustrate this point by considering the higher-order MSE of the oracle version of the DML1 and DML2 estimators defined in Remark 2.2. □

Remark 3.10. When K is fixed as $n \to \infty$, an explicit expression for the higher-order MSE of the oracle estimators $\hat{\theta}^*_{n,1}$ and $\hat{\theta}^*_{n,2}$ defined in Remark 2.2 can be derived based on standard arguments (e.g., Newey and Smith (2004)):

$$\begin{aligned}
\text{HO-MSE}[\hat{\theta}^*_{n,1}] &= \sigma^2/n + \left(K^2\Lambda^2 + K\Lambda_1\right)/n^2 + o(n^{-2}) \\
\text{HO-MSE}[\hat{\theta}^*_{n,2}] &= \sigma^2/n + \left(\Lambda^2 + \Lambda_1\right)/n^2 + o(n^{-2})
\end{aligned}$$

where

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

with σ^2 and Λ defined as in (2.11) and (3.6), respectively. Two main differences with respect to the results in Corollary 3.1 deserve further discussion. First, the remainder errors for the oracle versions are $o(n^{-2})$ and the second leading terms have a convergence rate of order n^{-2},

implying these terms are smaller than the second leading term of the higher-order MSE of $\hat{\theta}_{n,2}$ that is of order $n^{-1/2-2\varphi}$. Therefore, the second leading term that appears for the oracle estimators is a higher-order term included in $o(n^{-1/2-2\varphi})$. Second, the second leading term, $(K^2\Lambda + K\Lambda_1)/n^2$, in the higher-order MSE of $\hat{\theta}^*_{n,1}$ depends on K; therefore, for large values of K the accuracy of $\hat{\theta}^*_{n,1}$ is worse than the accuracy of $\hat{\theta}^*_{n,2}$. □

4 Lessons for Practitioners

This section presents some lessons for practitioners when implementing DML based on the formal results of Section 3. These lessons have theoretical support and, to the best of my knowledge, are new in the DML literature. Furthermore, these lessons are possible because of the asymptotic framework considered in this paper, which provides insights not captured by the existing first-order asymptotic theory or simulation-based evidence.

Before presenting them, it is important to remember that DML provides estimators as good as if the true nuisance function $\eta_0(\cdot)$ has been used. However, it gives a wide array of alternatives to practitioners that may seem roughly equivalent. Among these alternatives, two available estimators, namely, DML1 and DML2, are presented in Section 2.1. Each estimator depends on the number of equal-sized folds K in which the data are randomly split. In what follows, I present and discuss the recommendations for DML implementation, including how to select K.

First lesson: DML2 is the recommended option for DML implementation, especially in small-sample situations when increasing the number of folds is desired to improve the precision of the estimators $\hat{\eta}_k(\cdot)$, which use a fraction $(K-1)/K$ of the sample size n. This recommendation is not new. It was presented in Chernozhukov et al. (2018), but it now has a theoretical justification in terms of bias and MSE. The results in Section 3.2 show that the asymptotic distribution of DML2 is insensitive in terms of bias and MSE to the K values, which is not the case of DML1, which becomes increasingly sensitive in terms of bias and MSE to large values of K whenever the discrepancy measure Λ —defined in (3.6)— deviates from zero. Moreover, the conditions presented in Section 3.1 guarantee that the estimators based on DML2 are asymptotically valid for any $K \leq n$, including the leave-one-out estimator defined as DML2 with $K = n$.

Number of Folds for Cross-Fitting

The previous lesson recommends the use of DML2 estimators, but how to choose the number of folds is unclear because the results of Section 3.2 show that all DML2 estimators share the same asymptotic distribution. This question is addressed in what follows by considering two

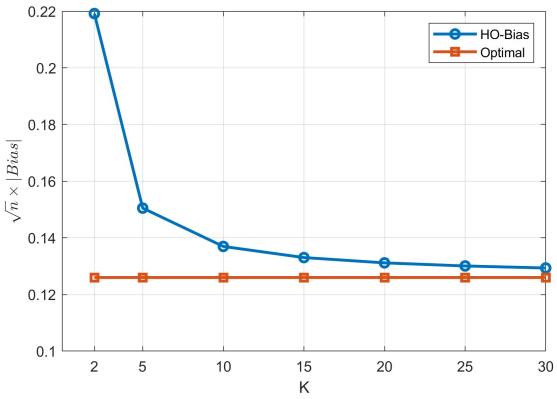


Figure 1: Higher-order bias for DML2, where $n = 1,000$, $F = 1$, and $\varphi = 2/5$.

different criterion based on the (higher-order) asymptotic bias and the second-order MSE. I used the explicit formulas presented in Section 3.3, where the higher-order properties of DML2 estimators were studied when K increases with the sample size n.

Second lesson: Choosing the number of folds equal to the sample size to implement DML2 is asymptotically optimal to reduce the (higher-order) asymptotic bias. The explicit formulas presented in Section 3.3 show that the absolute value of the leading term of the higher-order asymptotic bias for DML2 estimators is decreasing on K. For convenience, this explicit formula is presented next,

$$F\left(1+\frac{1}{K-1}\right)^{2\varphi} n^{-2\varphi} \tag{4.1}$$

where $\varphi \in (1/4, 1/2)$ and $F = |F_\delta + F_b|$. Figure 1 presents this explicit formula scaled by $\sqrt{n}$ as a function of the number of folds K that appears as a blue line with circular markers and the optimal level (minimal) obtained when $K = n$ that appears as a constant red line with square markers. Figure 1 uses $n = 1,000$, $F = 1$ and $\varphi = 2/5$ for illustrative purposes.

Therefore, choosing $K = n$ for DML2 implementation is optimal for reducing the asymptotic bias, while the common recommendations of choosing $K = 5, 10$ or 20 are suboptimal. Figure 1 reveals that the discrepancy between the asymptotic bias with the common recommendations for K and the optimal choice can be small. However, this conclusion may depend on the values of n, F, and φ used in Figure 1. In the fourth lesson, I discuss the relative loss of the asymptotic bias with respect to the optimal choice for the common choice of K for the arbitrary values of F and φ.

Third lesson: Choosing the number of folds equal to the sample size to implement DML2 can be asymptotically optimal for reducing the second-order asymptotic MSE. In other words, the leave-one-estimator defined as the DML2 estimator using $K = n$ can be the

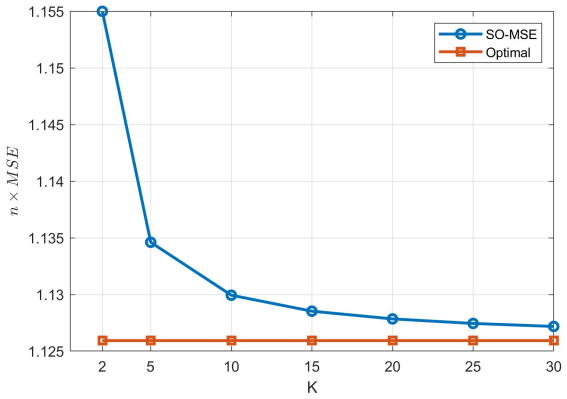


Figure 2: Second-order MSE for DML2, where $n = 1,000$, $\sigma = 1$, $G_b = 1$, and $\varphi = 2/5$.

most asymptotically accurate estimator among the class of DML2 estimators when a certain data-dependent condition holds.

To illustrate this lesson, I present next the second-order asymptotic MSE when the variance and the bias of the nuisance parameter estimator have the same convergence rate (i.e., $\varphi = \varphi_1 = \varphi_2$), that is,

$$\text{SO-MSE of } \hat{\theta}_{n,2} = \sigma^2/n + G_b \left(1 + \frac{1}{K-1}\right)^{2\varphi - 1/2} / n^{1/2+2\varphi} , \tag{4.2}$$

where $\varphi \in (0, 1/2)$ and G_b is a complex object that depends on the data. Note that SO-MSE of $\hat{\theta}_{n,2}$ in (4.2) is a decreasing function of K whenever G_b is positive, which is the data-dependent condition mentioned above. Figure 2 presents the second-order MSE scaled by n as a function of the number of folds K that appears as a blue line with circular markers and the optimal second-order MSE obtained by $K = n$ (under the assumption that $G_b > 0$) that appears as a constant red line with square markers. Figure 2 uses $n = 1,000$, $\sigma = 1$, $G_b = 1$, and $\varphi = 2/5$ for illustrative purposes.

Therefore, choosing $K = n$ for DML2 implementation is optimal for reducing the second-order MSE, while the common recommendations of $K = 5, 10$, or 20 are suboptimal under this criterion. Finally, the relative loss of the second-order MSE with respect to the optimal choice looks small for $K \geq 10$. However, the general conclusion may depend on the values of parameters (n, σ, G_b, φ). In the fourth lesson, I discuss the relative loss of the second-order asymptotic MSE with respect to the optimal choice for the common choice of K for arbitrary values of the parameters used in Figure 2.

Remark 4.1. Figure 2, and more concretely, the explicit expression for the second-order MSE presented in (4.2), explains several of the findings obtained through simulations, such

as the relatively large accuracy gains by increasing K from 2 to 5 folds compared to increases K from 5 to 10 or 10 to 20. □

The previous two lessons revealed that the common recommendation of choosing 5, 10, or 20 folds for the cross-fitting procedure in DML (e.g., Ahrens et al. (2024a,b), Bach et al. (2022), and Bach et al. (2024)) is suboptimal in terms of (higher-order) bias and second-order MSE. The next lesson discusses the relative loss a practitioner can face by choosing $K = 5$, 10, 20 instead of the optimal choice $K = n$.

Fourth lesson: If the optimal choice in terms of bias and accuracy is $K = n$, then choosing $K = 10$ for implementing DML2 guarantees that the maximum relative loss with respect to the optimal choice in terms of bias and second-order MSE is approximately 10% and 5%, respectively. In other words, the practitioner implementing DML2 with $K = 10$ has an estimator with a (higher-order) bias that is at most 10% larger than that obtained by implementing DML2 with the optimal $K = n$. Similarly, the second-order MSE of the DML2 estimator with $K = 10$ is at most 5% larger than that obtained by implementing DML2 with the optimal $K = n$. In what follows, I explain in more detail these results and present simple expressions for calculating these maximum relative losses as a function of K, n, and φ.

The relative loss of the (higher-order) bias with respect to the optimal choice is presented next as a function of K,

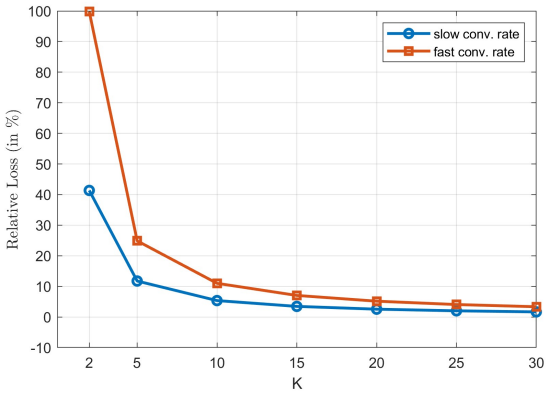
$$\left(\frac{1+\frac{1}{K-1}}{1+\frac{1}{n-1}}\right)^{2\varphi} - 1 \ , \tag{4.3}$$

which represents the percentage change of (4.1) with respect to the optimal value of (4.1) ($K = n$). In Figure 3, panel (a) presents the previous expression in percentages as a function of K. The blue line with circular markers represents the case of nuisance function estimators with slower convergence rates ($\varphi = 1/4$). The red line with square markers corresponds to nuisance function estimators with faster convergence rates ($\varphi = 1/2$). It follows from (4.3) and the figure that when $K = 10$, the relative loss of the (higher-order) bias is between 5% and 10% for different values of $\varphi \in (1/4, 1/2)$; therefore, it is at most 10%. Figure 3 consideres $n = 1,000$; nevertheless, the results do not change whenever $n \geq 1,000$.

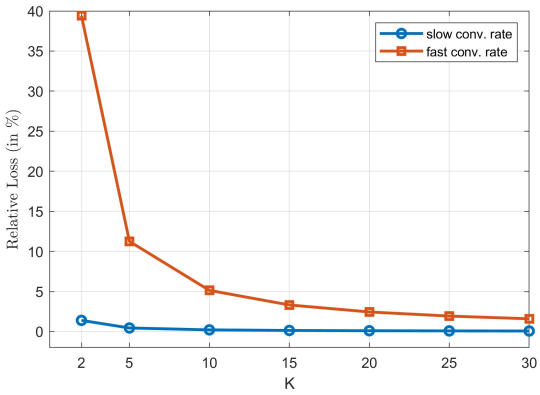
The relative loss of the second-order MSE with respect to the optimal choice is presented next as a function of K,

$$\frac{1+\upsilon\left(1+\frac{1}{K-1}\right)^{2\varphi-1/2}/n^{2\varphi-1/2}}{1+\upsilon\left(1+\frac{1}{n-1}\right)^{2\varphi-1/2}/n^{2\varphi-1/2}} - 1 \ , \tag{4.4}$$

which represents the percentage change of (4.2) with respect to the optimal value of (4.2)



(a) RL of Bias as in (4.3)



(b) RL of second-order MSE as in (4.5)

Figure 3: The relative loss (RL) of bias and second-order MSE with respect to the optimal choice for nuisance parameter estimators with slower ($\varphi = 1/4$) and faster ($\varphi = 1/2$) convergence rates, where $n = 1,000$.

when G_b is positive ($K = n$), and where $\upsilon = G_b/\sigma^2$. This previous expression depends on υ, which may be difficult to know in empirical applications. When υ is positive, the expression in (4.4) is an increasing function of υ for any $K \leq n$. In particular, (4.4) can be bounded by the following expression,

$$\frac{\left(1+\frac{1}{K-1}\right)^{2\varphi-1/2}}{\left(1+\frac{1}{n-1}\right)^{2\varphi-1/2}} - 1 \ , \tag{4.5}$$

which only depends on K, n, and $\zeta = 2\varphi - 1/2$. In Figure 3, panel (b) presents the previous expression in percentages as a function of K. It follows from (4.5) and the figure that when $K = 10$, the relative loss of the second-order asymptotic MSE is between 0.2% and 5%; therefore, it is at most 5%. Figure 3 consideres $n = 1,000$; nevertheless, the results do not change whenever $n \geq 1,000$.

Remark 4.2. The third lesson uses the second-order asymptotic MSE as an optimality criterion for evaluating the DML2 implementation. Specifically, it was used to guide how to select the number of folds K by minimizing the second-order MSE of $\hat{\theta}_{n,2}$. This criterion can be interpreted as an accuracy criterion because SO-MSE$[\hat{\theta}_{n,2}]$ can be interpreted as an approximation to the MSE of $\hat{\theta}_{n,2}$ based on the following informal derivations:

$$E\left[\left(\hat{\theta}_{n,2}-\theta_0\right)^2\right] \overset{(1)}{\approx} n^{-1}E\left[\mathcal{T}_{n,K}^2\right] \overset{(2)}{\approx} \text{SO-MSE}[\hat{\theta}_{n,2}] \ ,$$

where (1) is motivated by the stochastic expansion in Theorem 3.3 and (2) by the calculations based on Corollary 3.1. A similar criterion was used by Linton (1995) to select an optimal bandwidth and by Donald and Newey (2001) to select the optimal number of instruments,

while Newey and Smith (2004) used a similar idea to compare estimators sharing the same asymptotic distribution. □

Remark 4.3. The simulation results presented in Section 5 are consistent with $G_b > 0$ in (4.2). However, testing whether or not G_b is positive is left for future research. □

Remark 4.4. The second-order MSE described in Remark 4.2 defines an optimality criterion that can be used to compare different decisions regarding the DML implementation. In this paper, I used this criterion to select K. However, it can also be applied to select bandwidths or estimators for the nuisance function in applications. These alternative uses are beyond the scope of this paper and are left for future research. □

5 Monte-Carlo Simulations

This section examines the asymptotic results presented in Section 3 in finite samples. Specifically, I present the bias and mean squared error (MSE) of estimators based on DML1 and DML2 building on the Monte Carlo simulation designs for ATT-DID from Sant'Anna and Zhao (2020) and for LATE from Hong and Nekipelov (2010). Additionally, I present the coverage probability of confidence intervals constructed as in Chernozhukov et al. (2018). For the sake of readability, these confidence intervals are defined next

$$CI_j(1-\alpha) = \left[\hat{\theta}_{n,j} - z_{1-\alpha/2}\frac{\hat{\sigma}_{n,j}}{\sqrt{n}},\ \hat{\theta}_{n,j} + z_{1-\alpha/2}\frac{\hat{\sigma}_{n,j}}{\sqrt{n}}\right] , \tag{5.1}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution,

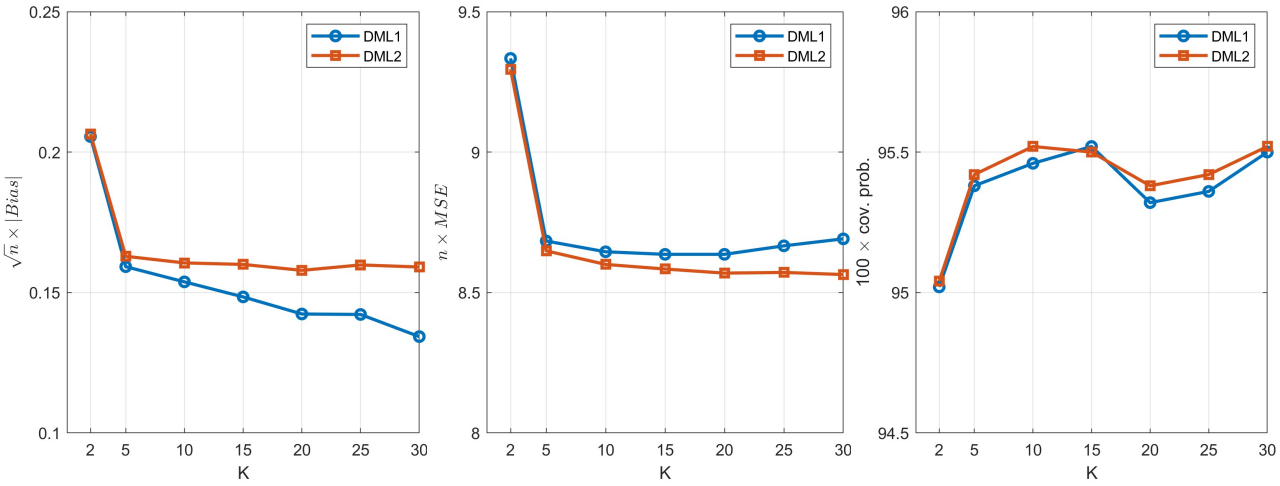
$$\hat{\sigma}^2_{n,j} = \frac{n^{-1}\sum_{i=1}^n m(W_i, \hat{\theta}_{n,j}, \hat{\eta}_i)^2}{\left(n^{-1}\sum_{i=1}^n \psi^a(W_i, \hat{\eta}_i)\right)^2} ,$$

$\hat{\theta}_{n,j}$ is as in (2.6) and (2.7) for DML1 and DML2, respectively, and $\hat{\eta}_i$ is as in (2.5).

5.1 Difference-in-Difference

This section is based on Example 2.2. I 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

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



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

Figure 4: Bias and MSE of estimators for the ATT-DID based on DML1 and DML2 as in (2.6) and (2.7), respectively. Coverage probability of confidence intervals as in (5.1) for the ATT-DID with a nominal level of 95%. Discrepancy measure $\Lambda = 0$, sample size $n = 3,000$ and 5,000 simulations.

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 $N(0, I_4)$, I_4 is the 4×4 identity matrix. The treatment assignment is defined by $A_i \sim \text{Bernoulli}(p(X_i))$, where

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

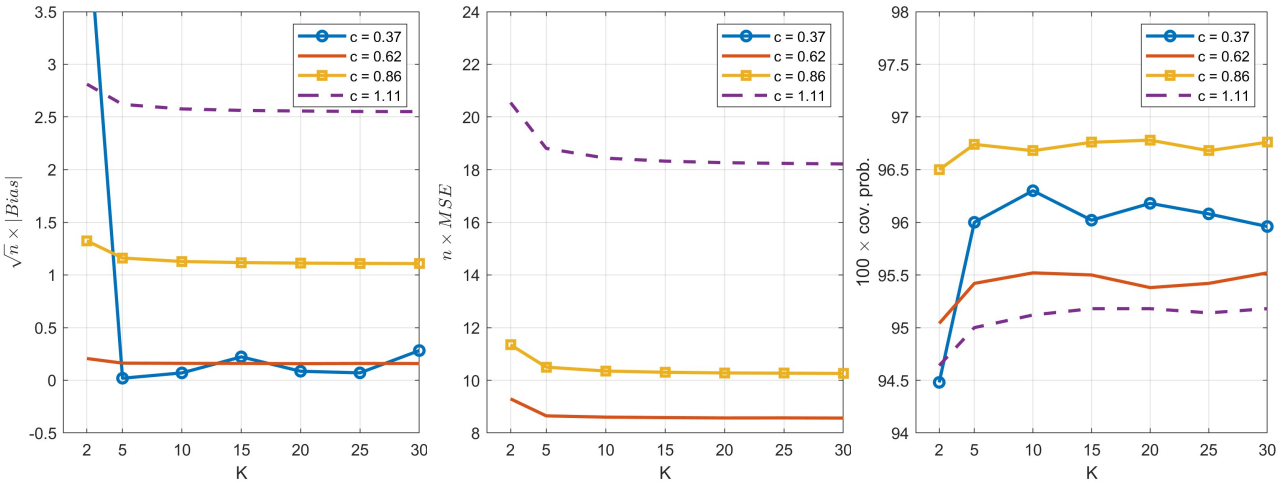
Finally, the vector of covariates is $X_i = (X_{1,i}, X_{2,i}, X_{3,i}, X_{4,i}) \in [0,1]^4$ and all its coordinates are independent uniform random variables (e.g., $X_{1,i} \sim \text{Uniform}[0,1]$).

The estimators for the ATT-DID are defined as in (2.6) and (2.7) using ψ^a and ψ^b presented in Example 2.2. To estimate the jth component of the vector of nuisance functions η_0, I use the Nadaraya-Watson estimator with a 6th-order Gaussian kernel and common bandwidth $h_j = cn_0^{-1/16}$ for all coordinates, where $n_0 = (K-1)/Kn$.[4]

I consider a sample size $n = 3,000$, different values for the choice of the number of folds $K \in \{2, 5, 10, \ldots, 30\}$, and perform 5,000 simulations. I additionally consider different values for the constant $c \in \{0.37, 0.62, 0.86, 1.11\}$ in the bandwidth.

Figures 4 and 5 report the results of the simulations in terms of bias, MSE, and coverage

[4]I also considered a 2nd order Gaussian Kernel in the simulations. The results are presented in Figures D.8 and D.9 in Appendix D, and they are similar to the ones presented using a 6th order Gaussian kernel.



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

Figure 5: Bias and MSE of estimators for the ATT-DID based on DML2 as in (2.7) 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 confidence intervals as in (5.1) for the ATT-DID with a nominal level of 95%. Sample size $n = 3,000$ and 5,000 simulations.

probability. Figure 4 compares the performance between the DML1 and DML2 estimators across different values of K, with $c = 0.62$. Panel (a) presents the absolute value of the scaled bias. It shows that the biases of DML1 and DML2 are similar. This result is consistent with the findings of Section 3.2 since the discrepancy measure $\Lambda = 0$ for Example 2.2 (ATT-DID). Furthermore, the bias of DML2 decreases as K increases, consistent with Theorem 3.4. Panel (b) presents the scaled MSE. It shows that DML1 and DML2 exhibit similar values. Moreover, the scaled MSE of DML2 decreases as K increases, which aligns with Corollary 3.1. Finally, Panel (c) highlights the similarities between DML1 and DML2 in terms of the coverage probability of their confidence intervals. Overall, Figure 4 aligns with the findings in Section 3, suggesting that DML1 and DML2 behave similarly when $\Lambda = 0$.

Figure 5 compares the results of DML2 estimators for different values of c. It presents results qualitatively similar to the ones in Figure 4, with some exceptions for $c = 0.37$ that present a non-monotonic scaled bias and large values for the scaled MSE. It also reveals that the scaled bias and MSE are sensitive to the values of c. For instance, the scaled MSE for $c = 1.11$ is larger than twice the scaled MSE for $c = 0.62$ due to a larger bias.

5.2 Local Average Treatment Effects

This section is based on Example 2.3. I 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

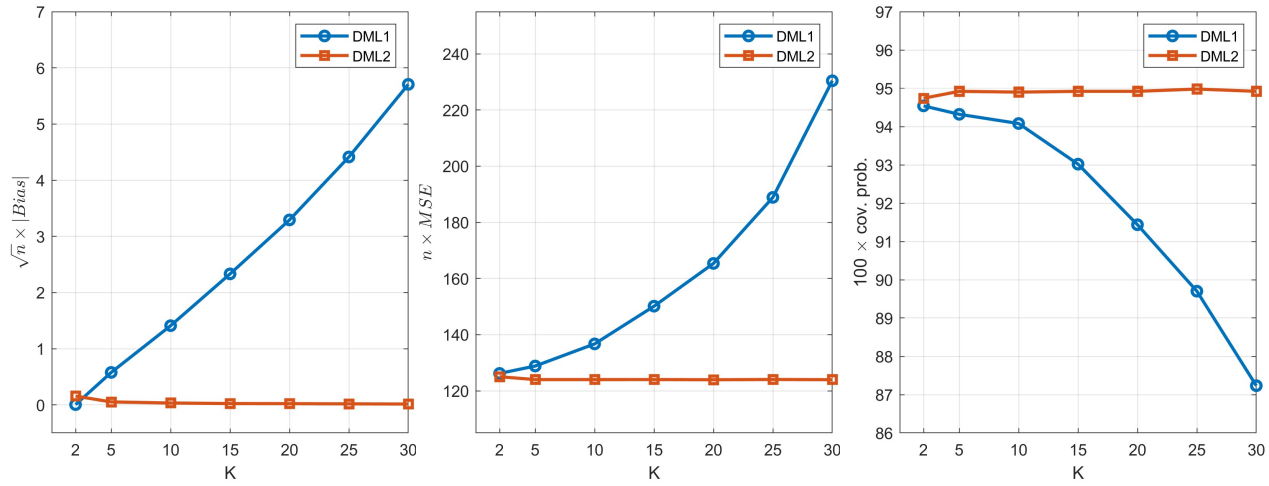
$$Y_i(1) = \xi_{1,i} + \xi_{3,i} I\{D_i(1) = 1, D_i(0) = 1\} + \xi_{4,i} I\{D_i(1) = 0, D_i(0) = 0\} \ ,$$
$$Y_i(0) = \xi_{2,i} + \xi_{3,i} I\{D_i(1) = 1, D_i(0) = 1\} + \xi_{4,i} I\{D_i(1) = 0, D_i(0) = 0\} \ ,$$

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 2.3, 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.

The estimators for the LATE are defined as in (2.6) and (2.7) using ψ^a and ψ^b presented in Example 2.3. To estimate the jth component of the vector of nuisance functions η_0, I use the Nadaraya-Watson estimator with a 2th-order Gaussian kernel and common bandwidth $h_j = c n_0^{-1/5}$, where $n_0 = (K-1)/Kn$.

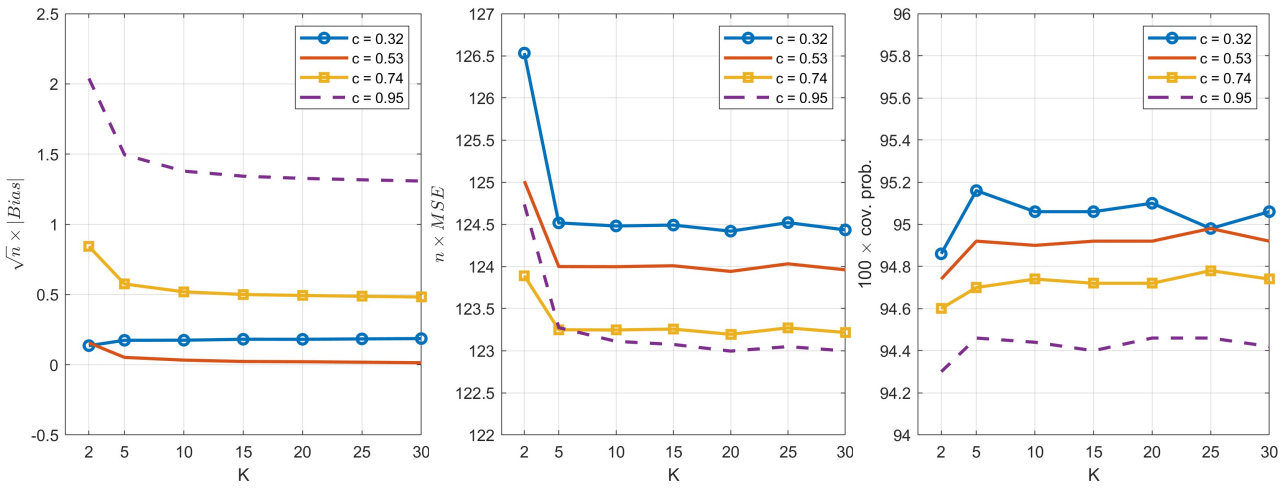
I consider a sample size $n = 3,000$, different values for the choice of the number of folds $K \in \{2, 5, 10, \ldots, 30\}$, and perform $5,000$ simulations. I additionally consider different values for the constant $c \in \{0.32, 0.53, 0.74, 0.95\}$ in the bandwidth.

Figures 6 and 7 present the results of the simulations in terms of bias, MSE, and coverage probability. Figure 6 compares the performance between the DML1 and DML2 estimators across different values of K, with $c = 0.53$. Panel (a) presents the absolute value of the scaled bias. It shows that the bias of DML1 is approximately an increasing linear function of K, which is consistent with the intuition presented in Remark 3.6 since the discrepancy measure Λ for the LATE in Example 2.3 is different than zero. In contrast, the bias of DML2 decreases as K increases, consistent with Theorem 3.4. Panel (b) presents the scaled MSE. It reveals that the scaled MSE for DML1 is increasing and approximately quadratic on K. This finding aligns with the expressions presented in Remark 3.10 for the oracle version of DML1. Additional simulation results presented in Figure D.10 in Appendix D show that the DML1 estimator and its oracle version exhibit similar values. In contrast, the scaled MSE



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

Figure 6: Bias and MSE of estimators for the LATE based on DML1 and DML2 as in (2.6) and (2.7), respectively. Coverage probability of confidence intervals as in (5.1) for the LATE with a nominal level of 95%. Discrepancy measure $\Lambda \neq 0$, sample size $n = 3,000$ and $5,000$ simulations.



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

Figure 7: Bias and MSE of estimators for the LATE based on DML2 as in (2.7) for different values of c in $h = cn_0^{-1/5}$. Coverage probability of confidence intervals as in (5.1) for the LATE with a nominal level of 95%. Sample size $n = 3,000$ and 5,000 simulations.

of DML2 is a decreasing function of K. Finally, Panel (c) shows dramatic discrepancies between DML1 and DML2 in terms of the coverage probability of the confidence intervals associated with them. In particular, it evidences that inference based on DML1 deteriorates as K increases. In contrast, DML2 does not have this problem, and inference is reliable for

all the values of K. Overall, Figure 6 is consistent with the findings in Section 3, suggesting that DML1 and DML2 behave differently when $\Lambda \neq 0$.

Figure 7 compares the results of DML2 estimators for different values of c. It presents results qualitatively similar to the ones in Figure 6, with some exceptions for $c = 0.32$ that present a non-monotonic scaled bias. It also reveals that the scaled bias and MSE are less sensitive to the values of c. For instance, the scaled MSE exhibits values between 123 and 124.5 for all the values of c and K between 5 and 30.

6 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 used to estimate the parameters of interest. In this paper, I consider an asymptotic framework in which K diverges to infinity as n diverges to infinity, accommodating small-sample situations where the practitioner may wish to increase K, situations not well approximated by the existing framework in which K is fixed as n diverges.

This paper makes several contributions within this new framework. First, it shows that DML2 asymptotically outperforms DML1 in terms of bias and mean squared error. Additionally, it characterizes the first-order asymptotic difference between DML1 and DML2 using a discrepancy measure, Λ, which can be calculated for various treatment effect parameters. 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, the paper employs higher-order asymptotic approximations that lead to the following final contribution: setting $K = n$ for DML2 implementation can be asymptotically optimal in terms of higher-order asymptotic bias and second-order asymptotic MSE within the class of DML2 estimators.

References

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*.

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

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

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.

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.

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.

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

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.

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.

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.

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.

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*.

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.

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.

A Additional Examples and Results

A.1 More Examples

Example A.1 (Weighted Average Treatment Effect)**.** This example is built on the setup of Example 2.1 (ATE) and the parameter considered in Equation (2) of Hirano et al. (2003). The parameter of interest is defined by

$$\theta_0 = E\left[E[(Y(1) - Y(0)) \mid X]g(X)\right] / E[g(X)] \ ,$$

where $g(\cdot)$ is a known function of covariates X, such that $|g(X)|$ is bounded and $E[g(X)] > 0$. When $g(X)$ equals the propensity score $E[A \mid X]$, the parameter θ_0 equals the average treatment effect on the treated, which implicitly assumes perfect knowledge of the propensity score. Under the selection-on-observables assumptions, the parameter θ_0 can be identified by a moment condition, such as (2.1), using a moment function like (2.2), where

$$\begin{aligned}
\psi^b(W, \eta) &= g(X)\left(\eta_1 - \eta_2 + A(Y - \eta_1)\eta_3 - (1 - A)(Y - \eta_2)\eta_4\right) \ , \\
\psi^a(W, \eta) &= g(X) \ ,
\end{aligned}$$

for $\eta \in \mathbf{R}^4$, and where the nuisance parameter $\eta_0(X)$ is exactly the same as in Example 2.1. This moment function appears as the efficient influence function in Hirano et al. (2003) for the weighted average treatment effect. In this example,

$$\Lambda = E[g(X)^2\left\{\eta_{0,1}(X) - \eta_{0,2}(X) - \theta_0\right\}]/E[g(X)]^2 \ ,$$

which is typically different than zero. □

Example A.2 (Average Treatment Effect on the Treated)**.** This example is built on the setup of Example 2.1 (ATE). It is assumed that there is no knowledge of the propensity score $E[A \mid X]$, and it has to be estimated. The parameter of interest is

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

which is the treatment effect for the treated group, also known as ATT. Under selection-on-observable assumptions, the parameter θ_0 can be identified by a moment condition, such as (2.1), using a moment function like (2.2), where

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

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

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

When there is no knowledge of the propensity score, this moment function appears as the efficient influence function for the ATT in Hahn (1998) and Hirano et al. (2003). In this example, $\Lambda = 0$. □

Example A.3 (Partial Linear Model)**.** This example presents the model studied in Robinson (1988) and Linton (1995). Consider the following model:

$$Y = D\theta_0 + g(X) + U \ ,$$

where $E[U \mid D, X] = 0$. Here $W = (Y, D, X)$. The parameter of interest is θ_0. In this example, θ_0 can be identified by (2.1) and (2.2), where

$$\psi^a(W, \eta) = (D - \eta_2)^2 \ ,$$
$$\psi^b(W, \eta) = (Y - \eta_1)(D - \eta_2) \ ,$$

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

$$\eta_{0,1}(X) = E[Y \mid X] \ ,$$
$$\eta_{0,2}(X) = E[D \mid X] \ .$$

In this example, $\Lambda = 0$. □

Example A.4 (Partial Linear IV Model)**.** This example presents the extended PLM. Consider the following model:

$$Y = D\theta_0 + g(X) + U \ ,$$
$$Z = m(X) + V \ ,$$

where $E[U \mid D, Z] = 0$ and $E[V \mid X] = 0$. Here $W = (Y, D, X, Z)$. The parameter of interest is θ_0. In this example, θ_0 can be identified by (2.1) and (2.2), where

$$\psi^a(W, \eta) = (D - \eta_2)(Z - \eta_3) \ ,$$
$$\psi^b(W, \eta) = (Y - \eta_1)(Z - \eta_3) \ ,$$

for $\eta \in \mathbf{R}^3$, and where the nuisance parameter $\eta_0(X)$ has three components:

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

In this example, Λ is typically different than zero. □

A.2 Additional Results for DML2 estimators

Let δ_{n_0} and b_{n_0} be the functions defined in Assumption 3.2. For $j \neq i$, define

$$\tilde{b}_{n_0}(X_i) = E\left[b_{n_0}(X_j, X_i) \mid X_i\right] \ .$$

Recall $\eta_i = \eta_0(X_i)$ and consider the following notation:

$$J_0 = E\left[\psi^a(W_i, \eta_i)\right] \ , \tag{A-1}$$

$$D_i = J_0^{-1}\left(\partial_\eta m(W_i, \theta_0, \eta)|_{\eta=\eta_i}\right) \ . \tag{A-2}$$

Using the previous notation, consider the stochastic approximation terms:

- $\mathcal{T}_{n,K}^l$ is the asymptotic second-order linear term and

$$\mathcal{T}_{n,K}^l = n^{-1/2} \sum_{i=1}^n \Delta_i^\top D_i \ , \tag{A-3}$$

  where Δ_i and D_i are as in (3.13) and (A-2), respectively.

- $\mathcal{T}_n^{dml2}$ is the asymptotic high-order DML term of the DML2 estimator and

$$\mathcal{T}_n^{dml2} = -n^{-1/2}\left(n^{-1/2}\sum_{i=1}^n m_i/J_0\right)\left(n^{-1/2}\sum_{i=1}^n(\psi_i^a - J_0)/J_0\right) \ . \tag{A-4}$$

- $\mathcal{T}_{n,K}^{dml1}$ is the asymptotic high-order DML term of the DML1 estimator

$$\mathcal{T}_{n,K}^{dml1} = -n^{-1/2}\sum_{k=1}^K\left(n_k^{-1/2}\sum_{i\in\mathcal{I}_k} m_i/J_0\right)\left(n_k^{-1/2}\sum_{i\in\mathcal{I}_k}(\psi_i^a - J_0)/J_0\right) \ , \tag{A-5}$$

  where $m_i = m(W_i, \theta_0, \eta_i)$, $\psi_i^a = \psi^a(W_i, \eta_i)$, and $n_k = n/K$.

The next assumption complements Assumption 3.3 to derive valid stochastic expansions for DML2 estimators when $\varphi_1 < \varphi_2$.

Assumption A.1.

(a) The following limits exist and are finite,

$$G_\delta^l = \lim_{n_0 \to \infty} E\left[\left(\delta_{n_0}(W_j, X_i)^\top D_i\right)\left(\delta_{n_0}(W_i, X_j)^\top D_j + \delta_{n_0}(W_j, X_i)^\top D_i\right)\right] , \tag{A-6}$$

$$G_b^l = \lim_{n_0 \to \infty} E\left[(m_i/J_0)\,\tilde{b}_{n_0}(X_i)^\top D_i\right] . \tag{A-7}$$

(b) $G_\delta^l > 0$.

The next theorem is an extension of Theorem 3.3. It considers valid stochastic expansions for $(\varphi_1, \varphi_2) \in (1/4, 1/2) \times (1/4, 1)$. It uses the following notation:

- $\mathcal{R}_1 = \{(\varphi_1, \varphi_2) \in (1/4, 1/2) \times (1/4, 1) : \varphi_1 < 1/3 \text{ or } \varphi_2 < 1/2\}$.
- $\mathcal{R}_2 = \{(\varphi_1, \varphi_2) \in (1/4, 1/2) \times (1/4, 1) : \varphi_1 \geq 1/3, \varphi_2 \geq 1/2, (\varphi_1, \varphi_2) \notin \mathcal{R}_3\}$.
- $\mathcal{R}_3 = \{(\varphi_1, \varphi_2) \in (1/4, 1/2) \times (1/4, 1) : \varphi_1 \geq 3/8, \varphi_2 \geq 1/2, \varphi_1 + \varphi_2 \geq 1\}$.

Theorem A.1. *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume that K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi_1 \in (1/4, 1/2)$, and $\varphi_1 \leq \varphi_2$, then*

$$n^{1/2}(\hat{\theta}_{n,2} - \theta_0) = \mathcal{T}_{n,K} + R_{n,K} , \tag{A-8}$$

where $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, $\hat{\theta}_{n,2}$ is as in (2.7), and

- *Case 1: $\mathcal{T}_{n,K} = \mathcal{T}_n^* + \mathcal{T}_{n,K}^{nl}$ if $(\varphi_1, \varphi_2) \in \mathcal{R}_1$, where $\mathcal{T}_n^*$ is defined in (3.11) and $\mathcal{T}_n^* \xrightarrow{d} N(0, \sigma^2)$ with σ^2 defined in (2.11), $\mathcal{T}_{n,K}^{nl}$ is defined in (3.12) and satisfies (i) $\lim_{n\to\infty} \inf_{K \leq n} Var[n^{2\varphi_1 - 1}\mathcal{T}_{n,K}^{nl}] > 0$ and (ii) $\lim_{n\to\infty} \sup_{K \leq n} E[(n^{2\varphi_1 - 1}\mathcal{T}_{n,K}^{nl})^2] < \infty$.*

  For the next two cases, suppose in addition that Assumption A.1 holds.

- *Case 2: $\mathcal{T}_{n,K} = \mathcal{T}_n^* + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_{n,K}^{l}$ if $(\varphi_1, \varphi_2) \in \mathcal{R}_2$, where $\mathcal{T}_n^*$ and $\mathcal{T}_{n,K}^{nl}$ are defined as in Case 1, $\mathcal{T}_{n,K}^{l}$ is defined in (A-3) and satisfies (i) $\lim_{n\to\infty} \inf_{K \leq n} Var[n^{\varphi_1}\mathcal{T}_{n,K}^{l}] > 0$ and (ii) $\lim_{n\to\infty} \sup_{K \leq n} E[(n^{\varphi_1}\mathcal{T}_{n,K}^{l})^2] < \infty$.*

- *Case 3: $\mathcal{T}_{n,K} = \mathcal{T}_n^* + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_{n,K}^{l} + \mathcal{T}_n^{dml2}$ if $(\varphi_1, \varphi_2) \in \mathcal{R}_3$, where $\mathcal{T}_n^*$, $\mathcal{T}_{n,K}^{nl}$, $\mathcal{T}_{n,K}^{l}$ are defined as in Case 1 and 2, $\mathcal{T}_n^{dml2}$ is defined in (A-4) and $n^{1/2}\mathcal{T}_n^{dml2}$ has a non-degenerate limit distribution.*

with

$$\lim_{n\to\infty} \sup_{K\leq n} P\left(n^{\zeta}|R_{n,K}| > \epsilon\right) = 0 \ ,$$

for any given $\epsilon > 0$.

Remark A.1. When K is fixed as $n \to \infty$ and $(\varphi_1, \varphi_2) \in \mathcal{R}_3$, the following stochastic expansion can be derived for DML1 estimators,

$$n^{1/2}(\hat{\theta}_{n,1} - \theta_0) = \mathcal{T}_n^{*} + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_n^{dml1} + o_p(n^{-\zeta}) \ ,$$

where $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$ and $\mathcal{T}_n^{dml1}$ is defined in (A-5). Furthermore, the stochastic expansion presented for DML2 in Theorem A.1 for case 3 ($(\varphi_1, \varphi_2) \in \mathcal{R}_3$) is also valid when K is fixed as $n \to \infty$. For convenience, it is presented below

$$n^{1/2}(\hat{\theta}_{n,2} - \theta_0) = \mathcal{T}_n^{*} + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_{n,K}^{nl} + \mathcal{T}_n^{dml2} + o_p(n^{-\zeta}) \ .$$

These expressions show that, for the values of $(\varphi_1, \varphi_2) \in \mathcal{R}_3$, both stochastic expansions are only different in $\mathcal{T}_n^{dml1}$ and $\mathcal{T}_n^{dml2}$, which capture the effects of implementing DML. □

Theorem A.2. *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume that K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi_1 \in (1/4, 1/2)$, $\varphi_2 < 1$, and $\varphi_1 \leq \varphi_2$, then*

$$E[\mathcal{T}_{n,K}] = F_K n^{1/2-2\varphi_1} + \nu_{n,K} \ ,$$

where

$$F_K = \begin{cases} (F_\delta + F_b)\left(1 + \frac{1}{K-1}\right)^{2\varphi_1} & \text{if } \varphi_1 = \varphi_2 \ , \\ F_\delta \left(1 + \frac{1}{K-1}\right)^{2\varphi_1} & \text{if } \varphi_1 < \varphi_2 \ , \end{cases} \tag{A-9}$$

and

$$\sup_{K\leq n} |\nu_{n,K}| = o(n^{1/2-2\varphi_1}) \ ,$$

with $\mathcal{T}_{n,K}$ defined as in Theorem A.1, and F_δ and F_b defined as in (3.3) and (3.4), respectively.

Theorem A.3. *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume that K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi_1 \in (1/4, 1/2)$, $\varphi_2 < 1$, and $\varphi_1 \leq \varphi_2$, then*

$$Var[\mathcal{T}_{n,K}] = \sigma^2 + \Omega_K/n^{\zeta} + r_{n,K} \ ,$$

where

$$\Omega_K = \begin{cases} G_b \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 > \varphi_2 \ , \\ \left(G_\delta \frac{K^2-3K+3}{(K-1)^2} + G_b\right) \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 = \varphi_2 \ , \\ G_\delta \left(\frac{(K^2-K+3)}{(K-1)^2}\right) \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 < \varphi_2 \ , \end{cases} \tag{A-10}$$

and

$$\sup_{K \leq n} |r_{n,K}| = o(n^{-\zeta}) \ ,$$

with $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, $\mathcal{T}_{n,K}$ *defined as in Theorem A.1, and* σ^2, G_δ *and* G_b *defined as in* (2.11), (3.2), *and* (3.5), *respectively.*

Corollary A.1. *Suppose Assumptions 3.1, 3.2, and 3.3 hold. In addition, assume that* K *is such that* $K \leq n$ *and* $K \to \infty$ *as* $n \to \infty$. *If* $\varphi_1 \in (1/4, 1/2)$, $\varphi_2 < 1$, *and* $\varphi_1 \leq \varphi_2$, *then*

$$E[\mathcal{T}_{n,K}^2] = \sigma^2 + \tilde{\Omega}_K / n^{\zeta} + \tilde{r}_{n,K} \ ,$$

where

$$\tilde{\Omega}_K = \begin{cases} G_b \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 > \varphi_2 \ , \\ \left(G_\delta \frac{K^2-3K+3}{(K-1)^2} + G_b + F_\delta^2 \left(\frac{K}{K-1}\right)\right) \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 = \varphi_2 \ , \\ \left(G_\delta \frac{(K^2-3K+3)}{(K-1)^2} + F_\delta^2 \left(\frac{K}{K-1}\right)\right) \left(\frac{K}{K-1}\right)^{\zeta} & \text{if } 3\varphi_1 - 1/2 < \varphi_2 \ , \end{cases} \tag{A-11}$$

$$\sup_{K \leq n} |\tilde{r}_{n,K}| = O(n^{1/2-2\varphi_1}) \ ,$$

with $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, $\mathcal{T}_{n,K}$ *defined as in Theorem A.1, and* σ^2, F_δ, G_δ, *and* G_b *defined as in* (2.11), (3.3), (3.5), *and* (3.2), *respectively.*

Tuning of Nuisance Parameters

The second-order asymptotic MSE can be defined by

$$\text{SO-MSE}[\hat{\theta}_{n,2}] = \sigma^2/n + \tilde{\Omega}_K / n^{\zeta+1} \ , \tag{A-12}$$

where σ^2 is the variance of the asymptotic distribution of the estimator $\hat{\theta}_{n,2}$ defined in (2.11), $\tilde{\Omega}_K$ is the higher-order asymptotic MSE defined in (A-11), $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, φ_1 is such that $n^{-2\varphi_1}$ is the convergence rate of the variance of the nuisance parameter estimator, and φ_2 is such that $n^{-\varphi_2}$ is the convergence rate of the bias of the nuisance parameter estimator.

The explicit formulas for the second-order asymptotic MSE show that tuning the nuisance parameter estimators optimally may still be suboptimal for estimating θ_0. Available recommendations for tuning the nuisance parameter estimators often rely on minimizing an out-of-sample prediction error, which can be interpreted as recommendations that minimize the asymptotic integrated mean squared error and guarantee optimal convergence rates for the nuisance parameter estimators. However, it is unclear if these recommendations are optimal for estimating the parameter of interest. In particular, the optimal tuning of the estimators for the nuisance parameter η_0 alone may be suboptimal for the parameter of interest θ_0. To illustrate this point, I consider Example 2.1 (ATE) using the Nadaraya-Watson (N-W) estimators for the nuisance parameter η_0. More concretely, I first obtain the values of φ_1 and φ_2 associated with the N-W estimators, and I then use them in (A-12) to derive the optimal convergence rate of the bandwidth based on the criterion.

For the sake of exposition I consider the case where the covariate X is univariate (e.g., $d_x = 1$), and the N-W uses a second-order kernel; see Appendix A.3 for additional details for the general case. Let $h_j = c_j n_0^{-\varphi_0}$ be the bandwidth used for estimating the component $\eta_{0,j}$ of η_0, where c_j is a given positive constant and $n_0 = ((K-1)/K)n$ is the sample size used for the estimation. In this case, it can be shown that

$$\varphi_1 = (1-\varphi_0)/2 \quad \text{and} \quad \varphi_2 = 2\varphi_0 \ , \tag{A-13}$$

where $\varphi_0 \in [1/5, 1/2)$ to guarantee that $\varphi_1 \in (1/4, 1/2)$ and $\varphi_1 \leq \varphi_2$. Using this notation, it follows that

$$\zeta = \min\{1 - 2\varphi_0, 3\varphi_0/2\} \ . \tag{A-14}$$

which is lower or equal than $3/7$.

In this example, the optimal convergence rate of the second term in (A-12) is $O(n^{-10/7})$ since $\zeta \leq 3/7$. This convergence rate is achieved when the bandwidth has a convergence rate $n^{-2/7}$ (i.e., $\varphi_0 = 2/7$). Importantly, this convergence rate is different than the optimal convergence rate of the bandwidth for the N-W estimators, which is $n^{-1/5}$. Therefore, the optimal tuning of the estimators for the nuisance parameter η_0 alone is suboptimal for the parameter of interest in this case.

Remark A.2. When $d_x = 1$ and the N-W estimator uses a second order kernel, then (A-13) also holds for Examples 2 and 3, and the other examples in Appendix A.1. Therefore, the conclusions derived here also apply to these examples. Appendix A.3 presents additional details and further discussion on the derivation of a general version of (A-13) when covariates X have dimension d_x and the N-W estimator uses a kernel of order $s > d_x/2$. □

A.3 Results for Nadaraya-Watson Estimators

This section presents explicit expressions for φ_1, φ_2 , δ_n, and b_n when the nuisance parameter estimator $\hat{\eta}$ is the Nadaraya-Watson (N-W) estimator. Specifically, it considers the N-W estimators with a common bandwidth for all coordinates, where the kernel function is the product of the same univariate kernels:

$$K_h(x) = h^{-d_x} \prod_{\ell=1}^{d_x} K(x_\ell/h) \ , \quad x \in \mathbf{R}^{d_x}$$

where $h = C_h n^{-\varphi_0}$ is the bandwidth and $K(\cdot)$ is a bounded symmetric kernel of order s.

Let $\eta_{0,j}(x)$ be the jth component of the nuisance parameter η_0. Let $f(x)$ be the density function of the covariates X at x. In what follows, I consider three possible types for this component:

Type 1: Simple conditional expectation, $\eta_{0,j}(x) = E[Y \mid X = x]$. In this case,

$$\hat{\eta}_{0,j}(x) = \frac{\sum_{\ell=1}^n Y_\ell K_h(x - X_\ell)}{\sum_{\ell=1}^n K_h(x - X_\ell)} \ .$$

By matching the convergence rate of the variance of this estimator ($(nh^d)^{-1}$) with $n^{-2\varphi_1}$, it follows that:

$$\varphi_1 = (1 - d_x\varphi_0)/2 \ ,$$

and by matching the convergence rate of the bias of this estimator (h^s) with $n^{-\varphi_2}$, it follows that:

$$\varphi_2 = s\varphi_0 \ .$$

Finally, the jth component of the functions δ_n and b_n are given by

$$\begin{aligned}
\delta_{n,j}(W, x) &= C_h^{-d_x/2} h^{d_x/2} \left(Y - \eta_{0,j}(X)\right) K_h\left(x - X\right)/f(x) \ , \\
b_{n,j}(W, x) &= h^{-s} \left(\eta_{0,j}(X) - \eta_{0,j}(x)\right) K_h\left(x - X\right)/f(x) \ ,
\end{aligned}$$

where X is a sub-vector of $W = (Y, X)$. Note that the dependence of these functions on n is due to the definition of bandwidth $h = C_h n^{-\varphi_0}$.

Type 2: Conditional expectation of a sub-group, $\eta_{0,j}(x) = E[Y \mid A = 1, X = x]$. In this case,

$$\hat{\eta}_{0,j}(x) = \frac{\sum_{\ell=1}^n Y_\ell A_\ell K_h(x - X_\ell)}{\sum_{\ell=1}^n A_\ell K_h(x - X_\ell)} \ .$$

The same arguments presented for type 1 imply

$$\varphi_1 = (1 - d_x\varphi_0)/2 \ ,$$
$$\varphi_2 = s\varphi_0 \ .$$

Finally, the jth component of the functions δ_n and b_n are given by

$$\delta_{n,j}(W,x) = C_h^{-d_x/2} h^{d_x/2} \left\{ (YA - g_1(X)) - \eta_{0,j}(x) \left(A - g_2(X)\right) \right\} K_h\left(x - X\right)/f(x) \ ,$$
$$b_{n,j}(W,x) = h^{-s} \left\{ (g_1(X) - g_1(x)) - \eta_{0,j}(x) \left(g_2(X) - g_2(x)\right) \right\} K_h\left(x - X\right)/f(x) \ ,$$

where $g_1(x) = E[YA \mid X = x]$ and $g_2(x) = E[A \mid X = x]$, X is a sub-vector of $W = (Y, A, X)$.

Type 3: Inverse of propensity score, $\eta_{0,j}(x) = (E[A \mid X = x])^{-1}$. In this case,

$$\hat{\eta}_{0,j}(x) = \frac{\sum_{\ell=1}^n K_h(x - X_\ell)}{\sum_{\ell=1}^n A_\ell K_h(x - X_\ell)} \ .$$

Under standard assumptions, it can be shown that $\hat{\eta}_{0,j}$ has the same convergence rates for the variance and bias. Therefore, the arguments presented for type 1 also apply and imply,

$$\varphi_1 = (1 - d_x\varphi_0)/2 \ ,$$
$$\varphi_2 = s\varphi_0 \ .$$

Finally, the jth component of the functions δ_n and b_n are given by

$$\delta_{n,j}(W,x) = -C_h^{-d_x/2} h^{d_x/2} \left(\eta_{0,j}(x)\right)^2 \left(A - g_2(X)\right) K_h\left(x - X\right)/f(x)$$
$$b_{n,j}(W,x) = -h^{-s} \left(\eta_{0,j}(x)\right)^2 \left(g_2(X) - g_2(x)\right) K_h\left(x - X\right)/f(x)$$

where $g_2(x) = E[A \mid X = x]$ and X is sub-vector of $W = (A, X)$.

Remark A.3. Any component of the nuisance parameters considered in Examples 2.1, 2.2, 2.3, A.1, A.2, A.3, and A.4 is of Type 1, 2, or 3, with minor modification (e.g., $\eta_{0,4} = E[1 - A \mid X]$ in Example 2.1 is of type 3). □

B Proofs of Main Results

B.1 Proof of Theorems 3.1 and 3.2

The proof of these theorems relies on the following decomposition,

$$n^{1/2}\left(\hat{\theta}_{n,j}-\theta_0\right)=n^{1/2}\left(\hat{\theta}_{n,j}-\hat{\theta}^*_{n,j}\right)+n^{1/2}\left(\hat{\theta}^*_{n,j}-\theta_0\right)\ . \tag{B-1}$$

and three intermediate results. The first two results are Theorems C.1 and C.2 in Appendix C that imply $n^{1/2}\left(\hat{\theta}_{n,j}-\hat{\theta}^*_{n,j}\right)=o_p(1)$ for $j=1,2$, respectively. These intermediate results rely on part (c) and (d) of Assumption 3.2 to accommodate the challenging situation that arises in the proof due to $K\to\infty$ as $n\to\infty$. The third result is Proposition C.1 in Appendix C that calculates the asymptotic distribution of $n^{1/2}\left(\hat{\theta}^*_{n,j}-\theta_0\right)$ for $j=1,2$. These three results and (B-1) complete the proof of the theorem.

B.2 Proof of Theorem 3.3

This follows from Theorem A.1 when $\varphi_1=\varphi_2$.

B.3 Proof of Theorem 3.4

This follows from Theorem A.2 when $\varphi_1=\varphi_2$.

B.4 Proof of Theorem 3.5

This follows from Theorem A.3 when $\varphi_1=\varphi_2$.

B.5 Proof of Theorem A.1

The proof of (A-8) in this theorem relies on the following decomposition,

$$n^{1/2}\left(\hat{\theta}_{n,2}-\theta_0\right)=n^{1/2}\left(\hat{\theta}_{n,2}-\hat{\theta}^*_{n,2}\right)+n^{1/2}\left(\hat{\theta}^*_{n,2}-\theta_0\right)\ . \tag{B-2}$$

and two intermediate results. The first result is Theorem C.2 in Appendix C that implies

$$n^{1/2}\left(\hat{\theta}_{n,2}-\hat{\theta}^*_{n,2}\right)=\mathcal{T}^{nl}_{n,K}+\mathcal{T}^{l}_{n,K}+\hat{R}_{n,K}\ ,$$

where $n^{\zeta}\hat{R}_{n,K}$ converges to zero in probability uniformly on $K\to\infty$ as $n\to\infty$ (equivalently, $\lim_{n\to\infty}\sup_{K\leq n}P(n^{\zeta}|\hat{R}_{n,K}|>\epsilon)=0$ for any given $\epsilon>0$). For case 1, it follows that $n^{\zeta}\mathcal{T}^{l}_{n,K}$

converges to zero in probability uniformly on $K \to \infty$ as $n \to \infty$. Proposition C.4 implies (i) $\lim_{n\to\infty} \inf_{K\leq n} Var[n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K}] > 0$ and $\lim_{n\to\infty} \sup_{K\leq n} E[(n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K})^2] < \infty$. For case 2 and 3, under Assumption A.1, Proposition C.3 implies that (i) $\lim_{n\to\infty} \inf_{K\leq n} Var[n^{\varphi_1}\mathcal{T}^{l}_{n,K}] > 0$ and $\lim_{n\to\infty} \sup_{K\leq n} E[(n^{\varphi_1}\mathcal{T}^{l}_{n,K})^2] < \infty$.

The second result is Proposition C.2 in Appendix C that implies

$$n^{1/2}\left(\hat{\theta}^{*}_{n,2} - \theta_0\right) = \mathcal{T}^{*}_{n} + \mathcal{T}^{dml2}_{n} + O_p(n^{-1}) \ . \tag{B-3}$$

This expansion is independent of the number of folds K since the oracle version of DML2 defined in (2.9) does not depend on sample splitting. Furthermore, it is valid for a larger class of parameters identified by (2.1) and can be obtained by standard arguments (e.g., Newey and Smith (2004)). Central Limit Theorem implies $\mathcal{T}^{*}_{n} \to N(0, \sigma^2)$, and Proposition C.2 implies $n^{1/2}\mathcal{T}^{dml2}_{n}$ has a non-degenerate limit distribution. Finally, note that in case 1 and 2, $n^{\zeta}\mathcal{T}^{dml2}_{n}$ converges to zero in probability uniformly on $K \to \infty$ as $n \to \infty$ (since $\mathcal{T}^{dml2}_{n}$ does not depend on K). In case 3, the remainder error term in (B-3) scaled by n^{ζ}, $n^{\zeta}O_p(n^{-2})$ converges to zero in probability uniformly on $K \to \infty$ as $n \to \infty$, since equation (B-3) does not depend on K.

The proof of (A-8) is completed by adding the previous two results in their respective case.

B.6 Proof of Theorem A.2

The proof of this theorem considers the definition of $\mathcal{T}_{n,K}$ with more terms (case 3),

$$\mathcal{T}_{n,K} = \mathcal{T}^{*}_{n} + \mathcal{T}^{l}_{n,K} + \mathcal{T}^{l}_{n,K} + \mathcal{T}^{dml2}_{n} \ ,$$

and two intermediate results. The first result is Proposition C.3 that shows $E[\mathcal{T}^{l}_{n,K}] = 0$ and Proposition C.4 that shows

$$E[\mathcal{T}^{nl}_{n,K}] = F_\delta \left(\frac{K}{K-1}\right)^{2\varphi_1} n^{1/2-2\varphi_1} + F_b \left(\frac{K}{K-1}\right)^{2\varphi_2} n^{1/2-2\varphi_2} + \nu_{n,K} \ .$$

The calculation of these terms relies on the structure imposed by part (a) of Assumption 3.2, and the Neyman orthogonality condition implied by part (b) of Assumption 3.1. The second result is Proposition C.2 in Appendix C that shows $E[\mathcal{T}^{*}_{n}] = 0$ and $E[\mathcal{T}^{dml2}_{n}] = \Lambda n^{-1/2}$. These two results and the definition of $\mathcal{T}_{n,K}$ complete the proof. For $\mathcal{T}_{n,K}$ in case 1 or 2, the proof is analogous and uses that $E[\mathcal{T}^{dml2}_{n}] = \Lambda n^{-1/2}$ do not depend on K and is $o(n^{1/2-2\varphi_1})$.

B.7 Proof of Theorem A.3

In the proof of the next theorem, $x_{n,K} = o(1)$ denotes a real valued sequence $x_{n,K}$ converging to zero uniformly on $K \to \infty$ as $n \to \infty$ (equivalently, $\lim_{n\to\infty} \sup_{K\leq n} |x_{n,K}| = 0$).

The proof of this theorem considers the definition of $\mathcal{T}_{n,K}$ with more terms (case 3) since the other two cases follow similarly.

$$\begin{aligned}
\text{Var}[\mathcal{T}_{n,K}] &= \text{Var}[\mathcal{T}_n^* + \mathcal{T}_{n,K}^l + \mathcal{T}_{n,K}^l + \mathcal{T}_n^{dml2}] \ , \\
&= \text{Var}[\mathcal{T}_n^* + \mathcal{T}_{n,K}^l] + \text{Var}[\mathcal{T}_{n,K}^l + \mathcal{T}_n^{dml2}] + 2\text{Cov}(\mathcal{T}_n^* + \mathcal{T}_{n,K}^l, \mathcal{T}_{n,K}^l + \mathcal{T}_n^{dml2}) \\
&\overset{(1)}{=} \text{Var}[\mathcal{T}_n^* + \mathcal{T}_{n,K}^l] + \text{Var}[\mathcal{T}_n^{dml2}] + 2\text{Cov}(\mathcal{T}_n^*, \mathcal{T}_n^{dml2}) + n^{-\zeta}o(1) \ , \qquad \text{(B-4)}
\end{aligned}$$

where (1) holds by the auxiliary results presented in Appendix C. More concretely, Proposition C.3 implies $\sup_{K\leq n} \text{Var}[\mathcal{T}_{n,K}^l] = o(n^{-\zeta})$ and $\sup_{K\leq n} \text{Cov}(\mathcal{T}_{n,K}^l, \mathcal{T}_n^{dml2}) = o(n^{-\zeta})$, respectively, part 3 and 4 of Proposition C.4 imply $\sup_{K\leq n} \text{Cov}(\mathcal{T}_{n,K}^{nl}, \mathcal{T}_{n,K}^l + \mathcal{T}_n^{dml2}) = o(n^{-\zeta})$, and part 1 of Proposition C.5 implies $\sup_{K\leq n} \text{Cov}(\mathcal{T}_n^*, \mathcal{T}_{n,K}^l) = o(n^{-\zeta})$.

To complete the proof, I present below explicit expressions for each of the first three terms in (B-4). Part 4 of Proposition C.2 is used to compute the second and third terms in (B-4). It shows

$$\text{Var}[\mathcal{T}_n^{dml2}] + 2\text{Cov}(\mathcal{T}_n^*, \mathcal{T}_n^{dml2}) = \Lambda_1 n^{-1} + O(n^{-2}) \ .$$

These calculations are independent of the assumptions on the nuisance parameter estimators and the number of folds since the oracle version of DML2 does not depend on sample splitting.

Part 3 of Proposition C.2 calculates $\text{Var}[\mathcal{T}_n^*]$, part 2 of Proposition C.4 calculates $\text{Var}[\mathcal{T}_{n,K}^l]$, and part 2 of Proposition C.5 calculates $\text{Cov}(\mathcal{T}_n^*, \mathcal{T}_{n,K}^l)$. All these expressions together compute the first term in (B-4). That is

$$\text{Var}[\mathcal{T}_n^* + \mathcal{T}_{n,K}^l] = \sigma^2 + G_\delta \left(\frac{K^2 - 3K + 3}{(K-1)^2} \right) n_0^{1-4\varphi_1} + G_b n_0^{1/2-\varphi_1-\varphi_2} + o(n^{-\zeta}) \qquad \text{(B-5)}$$

These calculations all depend on the structure imposed by part (a) of Assumption 3.2.

Finally, there are three possible cases based on $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$. First, if $3\varphi_1 - 1/2 > \varphi_2$ it follows that $\zeta = \varphi_1 + \varphi_2 - 1/2 < 4\varphi_1 - 1$ and the second term in (B-5) is $o(n^{-\zeta})$. In this case, the third term in (B-5) equals $\Omega_K n^{-\zeta}$. Second, if $3\varphi_1 - 1/2 = \varphi_2$, then $\zeta = 4\varphi_1 - 1$. In this case, the sum of the second and third terms in (B-5) equals $\Omega_K n^{-\zeta}$. Third, if $3\varphi_1 - 1/2 < \varphi_2$, then $\zeta = 4\varphi_1 - 1 < \varphi_1 + \varphi_2 - 1/2$ and the third term in (B-5) is $o(n^{-\zeta})$. In this case, the second term in (B-5) equals $\Omega_K n^{-\zeta}$.

C Auxiliary Results

The next result guarantees that a first-order equivalence property holds for the estimators based on DML and their oracle versions, even when K grows with the sample size.

Theorem C.1. *Suppose Assumptions 3.1 and 3.2 hold. In addition, assume K is such that $K \leq n$, $K \to \infty$ and $K/\sqrt{n} \to c \in [0, \infty)$ as $n \to \infty$. If $\varphi_1 \leq 1/2$ and $1/4 < \min\{\varphi_1, \varphi_2\}$, then*

$$n^{1/2}\left(\hat{\theta}_{n,1} - \hat{\theta}^*_{n,1}\right) = o_p(1)$$

where $\hat{\theta}_{n,1}$ and $\hat{\theta}^*_{n,1}$ are as in (2.6) and (2.8), respectively.

Proof. See Section E.1 in Appendix E □

Theorem C.2. *Suppose Assumptions 3.1 and 3.2 hold. In addition, assume that K is such that $K \leq n$ and $K \to \infty$ as $n \to \infty$. If $\varphi_1 \in (1/4, 1/2)$, and $\varphi_1 \leq \varphi_2$, then*

$$n^{1/2}\left(\hat{\theta}_{n,2} - \hat{\theta}^*_{n,2}\right) = o_p(1) \ ,$$

where $\hat{\theta}_{n,2}$ and $\hat{\theta}^*_{n,2}$ are defined as (2.7) and (2.9), respectively. Furthermore, if Assumption 3.3 holds, then

$$n^{1/2}\left(\hat{\theta}_{n,2} - \hat{\theta}^*_{n,2}\right) = \mathcal{T}^l_{n,K} + \mathcal{T}^l_{n,K} + \hat{R}_{n,K}$$

where

$$\lim_{n\to\infty} \sup_{K\leq n} P(n^{\zeta}|\hat{R}_{n,K}| > \epsilon) = 0 \ ,$$

for any fixed $\epsilon > 0$, with $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, $\mathcal{T}^l_{n,K}$ defined as in (3.12) and satisfies (i) $\lim_{n\to\infty} \inf_{K\leq n} Var[n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K}] > 0$ and (ii) $\lim_{n\to\infty} \sup_{K\leq n} E[(n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K})^2] < \infty$, and $\mathcal{T}^l_{n,K}$ defined as in (A-3) and satisfies $\lim_{n\to\infty} \sup_{K\leq n} E[(n^{\varphi_1}\mathcal{T}^l_{n,K})^2] < \infty$.

Proof. See Section E.2 in Appendix E. □

The next proposition calculates the asymptotic distribution of the oracle version of the DML estimators defined in Remark 2.2.

Proposition C.1. *Suppose Assumption 3.1. In addition, assume K is such that $K \leq n$, $K \to \infty$ and $K/n^{\gamma} \to c \in [0, \infty)$ as $n \to \infty$. Then,*

1. $n^{1/2}\left(\hat{\theta}^*_{n,1} - \theta_0\right) \xrightarrow{d} N(c\Lambda, \sigma^2)$ *when* $\gamma = 1/2$,
2. $n^{1/2}\left(\hat{\theta}^*_{n,2} - \theta_0\right) \xrightarrow{d} N(0, \sigma^2)$ *when* $\gamma = 1$,

where $\hat{\theta}^*_{n,1}$, $\hat{\theta}^*_{n,2}$, σ^2, *and* Λ *are as in* (2.8), (2.9), (2.11), *and* (3.6), *respectively.*

Proof. See Section E.3 in Appendix E □

The next proposition presents a stochastic expansion for the oracle version of the DML2 estimator, which does not depend on sample splitting or the number of folds K.

Proposition C.2. *Suppose Assumption 3.1 holds. Then,*

1. $n^{1/2}\left(\hat{\theta}^*_{n,2} - \theta_0\right) = \mathcal{T}^*_n + \mathcal{T}^{dml2}_n + O_p(n^{-1})$ *and* $E[\mathcal{T}^{dml2}_n] = \Lambda n^{-1/2}$
2. $E[\mathcal{T}^*_n] = 0$ *and* $E[(\mathcal{T}^*_n)^2] = \sigma^2$
3. $Cov(\mathcal{T}^*_n, \mathcal{T}^{dml2}_n) = -\Xi_1 n^{-1}$ *and* $Var[\mathcal{T}^{dml2}_n] = (\sigma^2\sigma_a^2 + \Lambda^2)n^{-1} + O(n^{-2})$

where

$$\Xi_1 = E\left[(m(W_i, \theta_0, \eta_i)/J_0)^2 (\psi^a(W_i, \eta_i) - J_0)/J_0\right] \tag{C-1}$$

$$\sigma_a^2 = E\left[((\psi^a(W_i, \eta_i) - J_0)/J_0)^2\right] \tag{C-2}$$

with $J_0 = E[\psi^a(W_i, \eta_i]$, $\eta_i = \eta_0(X_i)$, *and* $\hat{\theta}^*_{n,2}$, σ^2, Λ, $\mathcal{T}^*_n$, *and* $\mathcal{T}^{dml2}_n$ *defined as in* (2.9), (2.11), (3.6), (3.11), *and* (A-4), *respectively.*

Proof. See Section E.4 in Appendix E. □

Proposition C.3. *Suppose Assumptions 3.1, 3.2, and A.1 hold. In addition, assume that* K *is such that* $K \leq n$ *and* $K \to \infty$ *as* $n \to \infty$. *If* $\varphi_1 \leq \varphi_2$, *then*

1. $E[\mathcal{T}^l_{n,K}] = 0$
2. $Var[\mathcal{T}^l_{n,K}] = G^l_\delta \left(\frac{K}{K-1}\right)^{2\varphi_1} n^{-2\varphi_1} + r^l_{n,K}$, *where* $\sup_{K \leq n} |r^l_{n,K}| = o(n^{-2\varphi_1})$.
3. $\left|Cov(\mathcal{T}^{dml2}_n, \mathcal{T}^l_{n,K})\right| = o(n^{-2\varphi_1})$.

where $\mathcal{T}^l_{n,K}$, $\mathcal{T}^{dml2}_n$, *and* G^l_δ *are as in* (A-3), (A-4), *and* (A-6), *respectively.*

Proof. See Section E.5 in Appendix E. □

Proposition C.4. *Suppose Assumptions 3.1, 3.2, and A.1 hold. In addition, assume that* K *is such that* $K \leq n$ *and* $K \to \infty$ *as* $n \to \infty$. *If* $\varphi_1 \leq \varphi_2$ *and* $\varphi_1 \in (1/4, 1/2)$, *then*

1. $E[\mathcal{T}^{nl}_{n,K}] = F_\delta \left(\frac{K}{K-1}\right)^{2\varphi_1} n^{1/2-2\varphi_1} + F_b \left(\frac{K}{K-1}\right)^{2\varphi_2} n^{1/2-2\varphi_2} + \nu^{nl}_{n,K}$, *where* $\sup_{K \leq n} |\nu_{n,K}| = o(n^{1/2-2\varphi_1})$.
2. $Var[\mathcal{T}^{nl}_{n,K}] = G_\delta \left(\frac{(K^2-3K+3)}{(K-1)^2}\right) \left(\frac{K}{K-1}\right)^{4\varphi_1-1} n^{1-4\varphi_1} + r^{nl}_{n,K}$, *where* $|r^{nl}_{n,K}| = o(n^{-\zeta})$.

3. $\sup_{K\leq n}\left|Cov(\mathcal{T}_n^{dml2},\mathcal{T}_{n,K}^{nl})\right| = o(n^{-\zeta})$

4. $\sup_{K\leq n}\left|Cov(\mathcal{T}_{n,K}^{l},\mathcal{T}_{n,K}^{nl})\right| = o(n^{-\zeta})$

where $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, *and* $\mathcal{T}_{n,K}^{l}$, $\mathcal{T}_{n,K}^{nl}$, $\mathcal{T}_n^{dml2}$, F_δ, F_b, *and* G_δ *are as in* (A-3), (3.12), (A-4), (3.3), (3.4), *and* (3.2), *respectively.*

Proof. See Section E.6 in Appendix E. □

Proposition C.5. *Suppose Assumptions 3.1, 3.2, and A.1 hold. In addition, assume that* K *is such that* $K \leq n$ *and* $K \to \infty$ *as* $n \to \infty$. *If* $\varphi_1 \in (1/4, 1/2)$, $\varphi_2 < 1$, *and* $\varphi_1 \leq \varphi_2$. *Then,*

1. $Cov(\mathcal{T}_n^*, \mathcal{T}_{n,K}^{l}) = G_b^l \left(\frac{K}{K-1}\right)^{\varphi_2} n^{-\varphi_2} + r_{n,K}^{cov,l}$, *where* $\sup_{K\leq n}|r_{n,K}^{cov,l}| = o(n^{-\varphi_2})$.

2. $Cov(\mathcal{T}_n^*, \mathcal{T}_{n,K}^{l}) = \frac{G_b}{2}\left(\frac{K}{K-1}\right)^{1/2-\varphi_1-\varphi_2} n^{1/2-\varphi_1-\varphi_2} + r_{n,K}^{cov,nl}$, *where* $\sup_{K\leq n}|r_{n,K}^{cov,nl}| = o(n^{-\zeta})$.

where $\zeta = \min\{4\varphi_1 - 1, \varphi_1 + \varphi_2 - 1/2\}$, *and* $\mathcal{T}_n^*$, $\mathcal{T}_{n,K}^{l}$, $\mathcal{T}_{n,K}^{nl}$, G_b, *and* G_b^l *are as in* (3.11), (A-3), (3.12), (3.5), *and* (A-7), *respectively.*

Proof. See Section E.7 in Appendix E. □

The next lemma is useful to prove intermediate results such as Lemmas C.2 and C.3.

Lemma C.1. *Let Assumption 3.2 hold. Then, there exists a positive constant* $C = C(p, M_1)$ *such that for any* $i \in \mathcal{I}_k$ *and* $k \in \{1, \ldots, K\}$

1. $E\left[\left|\left|n_0^{-1/2}\sum_{\ell\notin\mathcal{I}_k} n_0^{-\varphi_1}\delta_{n_0}(W_\ell, X_i)\right|\right|^4\right] \leq C n_0^{-4\varphi_1}$

2. $E\left[\left|\left|n_0^{-1}\sum_{\ell\notin\mathcal{I}_k} n_0^{-\varphi_2} b_{n_0}(W_\ell, X_i)\right|\right|^4\right] \leq C(n_0^{-4\varphi_1} + n_0^{-4\varphi_2})$

Proof. See Section E.8 in Appendix E. □

Lemma C.2. *Suppose that Assumptions 3.1 and 3.2 hold. In addition, assume that* K *is such that* $K \leq n$ *and* $K \to \infty$ *as* $n \to \infty$. *If* $\varphi_1 \leq 1/2$ *and* $1/4 < \min\{\varphi_1, \varphi_2\}$, *then*

1. *For* $z = a, b$,

$$\lim_{\tilde{M}\to\infty}\lim_{n\to\infty}\sup_{K\leq n} P\left(n^{\min\{\varphi_1,\varphi_2\}+1/2}\left|n^{-1}\sum_{i=1}^{n}(\hat{\eta}_i - \eta_i)^\top \partial_\eta \psi^z(W_i, \eta_i)\right| > \tilde{M}\right) = 0\ ,$$

2. *For* $z = a, b$,

$$\lim_{\tilde{M}\to\infty}\lim_{n\to\infty}\sup_{K\leq n} P\left(n^{2\min\{\varphi_1,\varphi_2\}}\left|n^{-1}\sum_{i=1}^{n}\psi^z(W_i, \hat{\eta}_i) - \psi^z(W_i, \eta_i)\right| > \tilde{M}\right) = 0\ ,$$

3.

$$\lim_{\tilde{M}\to\infty}\lim_{n\to\infty}\sup_{K\leq n} P\left(n^{\min\{\varphi_1,\varphi_2\}-1/2}\left|n^{-1}\sum_{i=1}^{n}(\hat{\eta}_i-\eta_i)^\top\partial_\eta m(W_i,\theta_0,\eta_i)\right|>\tilde{M}\right)=0\ ,$$

4. Let $Z_{n,K}=n^{-1/2}\sum_{i=1}^{n}\left(m(W_i,\theta_0,\hat{\eta}_i)-m(W_i,\theta_0,\eta_i)\right)/J_0-\left(\mathcal{T}^l_{n,K}+\mathcal{T}^{nl}_{n,K}\right)$, *then*

$$\lim_{\tilde{M}\to\infty}\lim_{n\to\infty}\sup_{K\leq n} P\left(n^{-1/2+3\min\{\varphi_1,\varphi_2\}}\left|Z_{n,K}\right|>\tilde{M}\right)=0\ ,$$

where $\eta_i=\eta_0(X_i)$, $\hat{\eta}_i$ *is as in* (2.5), $J_0=E[\psi^a(W_i,\eta_i)]$, *and* $\mathcal{T}^l_{n,K}$ *and* $\mathcal{T}^{nl}_{n,K}$ *are as in* (A-3) *and* (3.12), *respectively. Moreover, it holds* $\lim_{n\to\infty}\sup_{K\leq n}E[(n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K})^2]<\infty$ *and* $\lim_{n\to\infty}\sup_{K\leq n}E[(n^{\varphi_1}\mathcal{T}^{l}_{n,K})^2]<\infty$.

Furthermore, if Assumption 3.3 holds, then $\lim_{n\to\infty}\inf_{K\leq n}Var[n^{2\varphi_1-1}\mathcal{T}^{nl}_{n,K}]>0$; *and if Assumption A.1 holds, then* $\lim_{n\to\infty}\inf_{K\leq n}Var[n^{\varphi_1}\mathcal{T}^{l}_{n,K}]>0$.

Proof. See Section E.9 in Appendix E. □

Lemma C.3. *Suppose Assumptions 3.1 and 3.2 hold. In addition, assume* K *is such that* $K\leq n$, $K\to\infty$ *and* $K/\sqrt{n}\to c\in[0,+\infty)$ *as* $n\to\infty$. *If* $1/4<\min\{\varphi_1,\varphi_2\}$ *and* $\varphi_1\leq 1/2$, *then,*

$$\lim_{n\to\infty}\sup_{K\leq n} P\left(\max_{k=1,\ldots,K}\left|n_k^{-1/2}\sum_{i\in\mathcal{I}_k}(\hat{\eta}_i-\eta_i)^\top\partial_\eta\psi^z(W_i,\eta_i)\right|>\epsilon\right)=0\ , \tag{C-3}$$

and

$$\lim_{n\to\infty}\sup_{K\leq n} P\left(\max_{k=1,\ldots,K} n_k^{-1}\sum_{i\in\mathcal{I}_k}||\hat{\eta}_i-\eta_i||^2>\epsilon\right)=0\ , \tag{C-4}$$

for $z=a,b$, *where* $\hat{\eta}_i$ *is as in* (2.5) *and* $\eta_i=\eta_0(X_i)$. *In particular,*

$$\lim_{n\to\infty}\sup_{K\leq n} P\left(\max_{k=1,\ldots,K}\left|n_k^{-1/2}\sum_{i\in\mathcal{I}_K}(\hat{\eta}_i-\eta_i)^\top\partial_\eta m(W_i,\theta_0,\eta_i)\right|>\epsilon\right)=0\ , \tag{C-5}$$

and

$$\lim_{n\to\infty}\sup_{K\leq n} P\left(\max_{k=1,\ldots,K}\left|n_k^{-1}\sum_{i\in\mathcal{I}_k}\psi^a(W_i,\hat{\eta}_i)-\psi^a(W_i,\eta_i)\right|>\epsilon\right)=0 \tag{C-6}$$

for any given $\epsilon>0$.

Proof. See Section E.10 in Appendix E. □

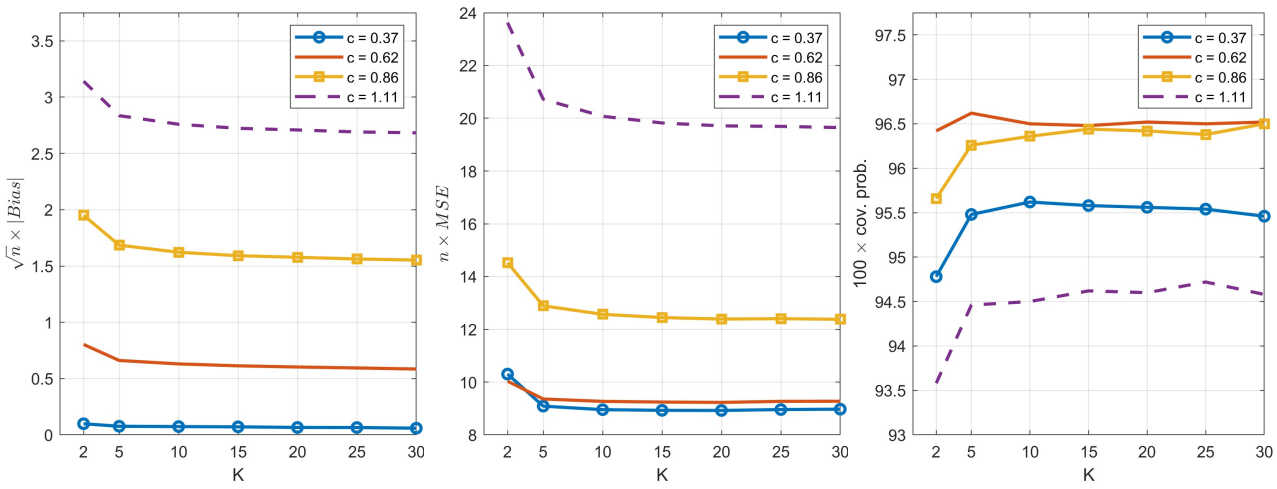
Lemma C.4. *Suppose Assumption 3.2 holds. In addition, assume* K *is such that* $K\leq n$, $K\to\infty$ *as* $n\to\infty$. *If* $\varphi_1\leq 1/2$, *then*

1. $\lim_{\tilde{M}\to\infty} \lim_{n\to\infty} \sup_{K\leq n} P\left(n^{4\min\{\varphi_1,\varphi_2\}}(n^{-1}\sum_{i=1}^{n}||\hat{\eta}_i-\eta_i||^4) > \tilde{M}\right) = 0$,
2. $\lim_{n\to\infty} \sup_{K\leq n} n^{2\min\{\varphi_1,\varphi_2\}} E\left[||\hat{\eta}_i-\eta_i||^2\right] < \infty$,
3. $\lim_{\tilde{M}\to\infty} \lim_{n\to\infty} \sup_{K\leq n} P\left(n^{2\min\{\varphi_1,\varphi_2\}} n^{-1}\sum_{i=1}^{n}||\hat{\eta}_i-\eta_i||^2 > \tilde{M}\right) = 0$,
4. $\lim_{n\to\infty} \sup_{K\leq n} P\left(n^{-1/2}\sum_{i=1}^{n}||\hat{\eta}_i-\eta_i||^2 > \epsilon\right) = 0$ *when* $1/4 < \min\{\varphi_1,\varphi_2\}$, *for any given* $\epsilon > 0$.

where $\eta_i = \eta_0(X_i)$ *and* $\hat{\eta}_i$ *is as in* (2.5).

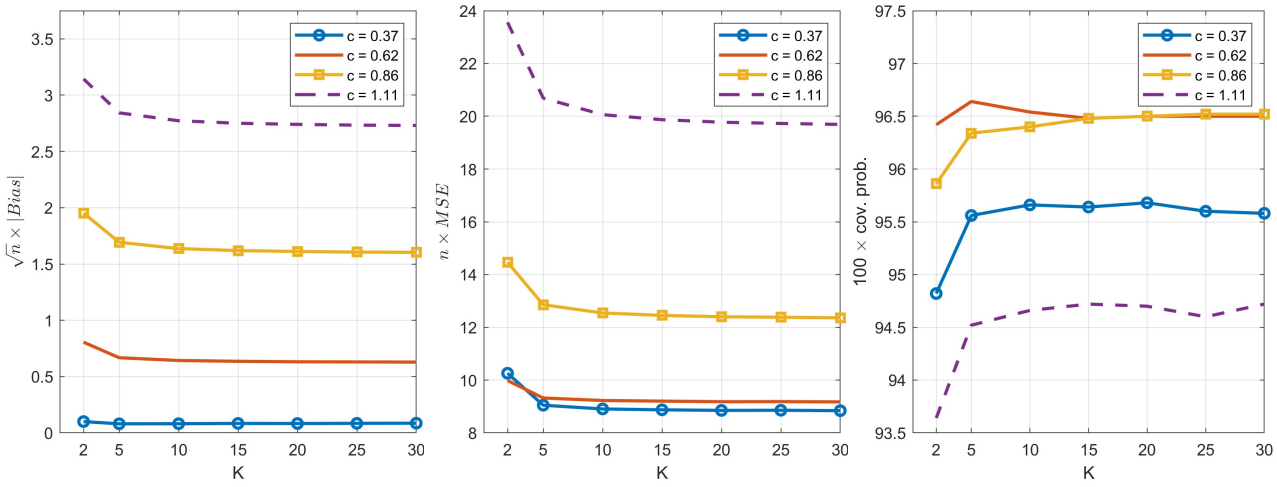
Proof. See Section E.11 in Appendix E. □

D Additional Simulation Results



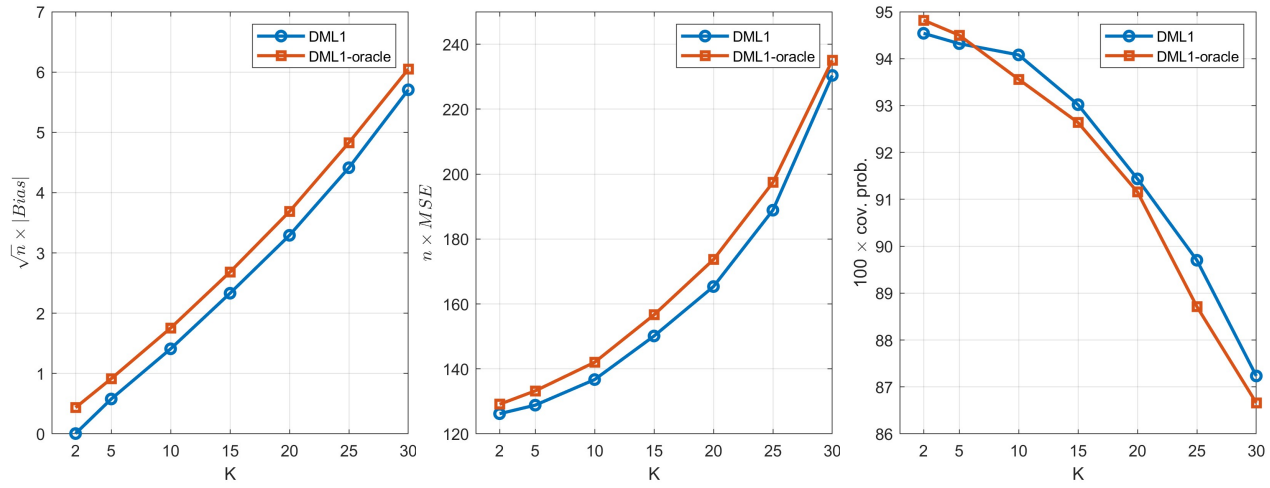
(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.8: Bias and MSE of estimators for the ATT-DID based on DML1 as in (2.7) for different values of c in $h = cn_0^{-1/5}$. Coverage probability of confidence intervals as in (5.1) for the ATT-DID with a nominal level of 95%. Sample size $n = 3,000$ and 5,000 simulations. It uses a Second Order Gaussian Kernel



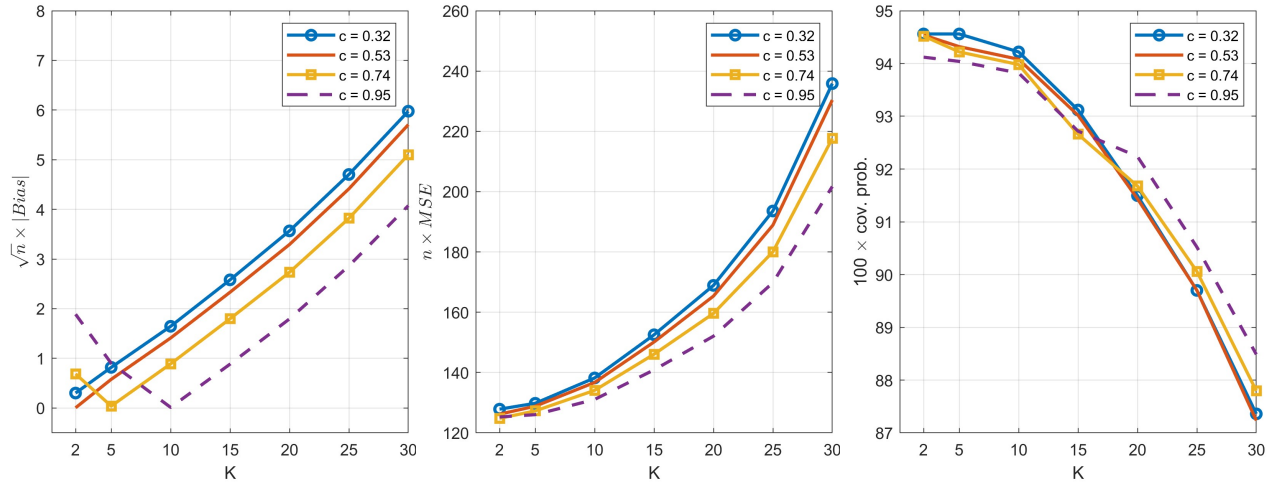
(a) Bias (b) MSE (c) Coverage probability (%)

Figure D.9: Bias and MSE of estimators for the ATT-DID based on DML2 as in (2.7) for different values of c in $h = cn_0^{-1/5}$. Coverage probability of confidence intervals as in (5.1) for the ATT-DID with a nominal level of 95%. Sample size $n = 3,000$ and 5,000 simulations. It uses a Second Order Gaussian Kernel



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

Figure D.10: Bias and MSE of estimators for the LATE based on DML1 and DML2 as in (2.8) and (2.9), respectively. Coverage probability of confidence intervals as in (5.1) but using true σ^2 for the LATE with a nominal level of 95%. Discrepancy measure $\Lambda \neq 0$, sample size $n = 3,000$ and $5,000$ simulations.



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

Figure D.11: Bias and MSE of estimators for the LATE based on DML1 as in (2.6) for different values of c in $h = cn_0^{-1/5}$. Coverage probability of confidence intervals as in (5.1) for the LATE with a nominal level of 95%. Discrepancy measure $\Lambda \neq 0$, sample size $n = 3,000$ and $5,000$.