

# On the Asymptotic Properties of Debiased Machine Learning Estimators \*

Amilcar Velez

Department of Economics, Cornell University

[amilcare@cornell.edu](mailto:amilcare@cornell.edu)

This version: March 20, 2026.

Newest version [here](#).

## Abstract

This paper studies debiased machine learning (DML) under a novel asymptotic framework. DML is a two-step estimation method for econometric models in which the parameter of interest depends on unknown nuisance functions. It uses  $K$ -fold cross-fitting to accommodate flexible machine-learning estimators. Practitioners implementing DML confront multiple decisions: whether to use DML1 or DML2 (two variants of DML estimators), and how to choose  $K$ . Existing fixed- $K$  asymptotic theory establishes that DML1 and DML2 are asymptotically equivalent, offering no formal guidance on which variant to use or how to select  $K$ . Under a framework in which  $K$  can grow with the sample size  $n$ , we demonstrate that DML2 offers theoretical advantages over DML1 in terms of bias, mean-squared error (MSE), and inference. When first-step estimators admit a linear stochastic expansion, we further show that for scalar DML2 the choice  $K = n$  is asymptotically optimal in terms of second-order asymptotic bias and MSE.

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

---

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

# 1 Introduction

Debiased machine learning (DML) is a two-step estimation method for econometric settings in which the parameter of interest depends on unknown nuisance functions (Chernozhukov et al., 2018). By combining  $K$ -fold cross-fitting with Neyman orthogonality, DML attains standard asymptotic properties under milder conditions than classical semiparametric methods, allowing machine-learning methods in the first-step. In practice, we use either DML1 or DML2 (two existing variants) and select  $K$ . Existing fixed- $K$  asymptotic theory establishes that both variants have identical limiting distributions, suggesting they are interchangeable and providing no guidance for choosing between them or for selecting  $K$ . However, simulation evidence shows that DML2 outperforms DML1 for some econometric models (Chernozhukov et al., 2018), and increasing  $K$  can reduce bias and mean squared error (MSE) of DML2 (Ahrens et al., 2025).

This paper studies the properties of DML1 and DML2 under a novel asymptotic framework in which  $K$  can grow with the sample size  $n$ . Under this framework, we demonstrate that DML2 is never worse and can be strictly better than DML1 in terms of bias, MSE, and inference. We characterize when this dominance is strict through a model-dependent quantity  $\Lambda$ . When  $\Lambda \neq 0$  and  $K \propto \sqrt{n}$ , the limiting distribution of DML1 has a nonzero asymptotic bias, making DML1-based inference invalid. In contrast, DML2 remains robust when  $K \propto \sqrt{n}$ , and continues to do so when  $K \propto n$  under additional assumptions. We then restrict the class of first-step estimators—to those admitting a linear stochastic expansion—to derive a second-order asymptotic approximation for scalar DML2 estimators, showing that larger  $K$  reduces DML2’s second-order bias and MSE, making  $K = n$  an asymptotically optimal choice.

DML is an estimation method applicable to econometric models in which the parameter of interest  $\theta_0$  is finite dimensional and satisfies a moment condition of the following form:

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

where  $m$  is a known moment function,  $W$  is an observed random vector, and  $\eta_0$  is an unknown nuisance function. Many examples of  $\theta_0$  satisfying (1.1) include the average treatment effect (ATE), average treatment effect on the treated in difference-in-differences designs (ATT-DID), local average treatment effect (LATE), coefficient in the partially linear model (PLM) and partially linear IV model (PLM-IV), among others studied in the literature (e.g., Robinson (1988), Robins et al. (1994), Hahn (1998), Hirano et al. (2003), Frölich (2007), and Sant’Anna and Zhao (2020)). In all these examples, the moment function  $m$  is linear in the parameter  $\theta_0$ , and the nuisance function  $\eta_0$  is based on conditional expectations, such as the

propensity score. This paper considers a setup that includes all these examples.

DML relies on two ingredients to attain standard asymptotic properties (e.g., asymptotic normality with parametric rates). The first is *Neyman orthogonality*, a condition on the moment function  $m$  that guarantees that the estimation of  $\theta_0$  is as accurate as if the true  $\eta_0$  were used; see Remarks 3.1 and 3.2. The second ingredient is *cross-fitting*, a form of sample splitting used to estimate the nuisance functions, which complements orthogonality by accommodating a large class of first-step estimators, including machine-learning methods.

Two variants of the DML estimator for  $\theta_0$  were proposed by Chernozhukov et al. (2018): DML1 and DML2. Both randomly divide the data into  $K$  equal-sized folds to estimate the nuisance functions, but they differ in how these estimates are combined. We provide a precise description of DML1 and DML2 in Section 2. Because each variant can be implemented with different values of  $K$ , practitioners face multiple decisions such as choosing between variants and selecting  $K$ . This paper aims to provide guidance on these choices.

Existing fixed- $K$  asymptotic theory states DML1 and DML2 are asymptotically equivalent for any  $K$ , offering no formal guidance on choosing between them or on the choice of  $K$ . Nevertheless, simulation evidence indicates that DML2 outperforms DML1 in some models, while in others their performance is similar. As a result, the literature recommends DML2 as a default, but theoretical reasons for this recommendation remain unknown. How to choose  $K$  for DML2 is also an open question; common practice is to use  $K = 5$  or  $10$ .

This paper studies the properties of DML1 and DML2 under a novel asymptotic framework in which  $K = K_n$  can grow with  $n$ , aiming to understand which version has theoretical advantages. Our framework provides a better approximation to finite-sample situations where practitioners want to use large  $K_n$  values, including  $K_n = n$ . Our approach follows a tradition in econometrics of using refined asymptotic approximations to study finite-sample behavior, as in Cattaneo and Jansson (2018), Bugni and Canay (2021), and Cai (2022).

Our first theoretical result in Section 3 explains when and why DML2 outperforms DML1, formalizing existing simulation evidence that fixed- $K$  asymptotic theory cannot explain. Specifically, we show that the limiting distribution of DML1 is sensitive to the choice of  $K_n$  when  $\Lambda$ —a model-based quantity defined in (3.1)—is different from zero. When  $K_n \propto \sqrt{n}$ , we demonstrate that the first-order asymptotic bias of DML1 is proportional to  $\Lambda$ , implying that inference based on DML1 can be invalid. In contrast, DML2 continues to be valid for  $K_n \propto \sqrt{n}$ , implying that inference based on DML2 remains valid.

We then consider an additional assumption under which DML2 yields valid inference for any choice of  $K_n$ , including  $K_n = n$ , which corresponds to the leave-one-out estimator. The assumption is an algorithm stability condition requiring that first-step estimators are stable to replacing a single observation with another i.i.d. draw—a condition we verify for kernel

estimators. Under this assumption, together with standard conditions used in the DML literature, we prove that the limiting distribution of DML2 is the same for any  $K_n$ .

Lastly, to provide guidance on the choice of  $K_n$ , we focus on scalar DML2 estimators and first-step estimators that admit a linear stochastic expansion (e.g., kernel estimators with MSE-optimal bandwidths). Under these conditions, we derive a valid second-order asymptotic approximation for DML2 and show that increasing  $K_n$  reduces the second-order asymptotic bias and MSE, implying that commonly recommended choices such as  $K_n = 5$  or 10 are suboptimal. Moreover, we use our approximation to quantify the relative efficiency loss in terms of second-order MSE from these suboptimal choices, and find that using  $K_n = 10$  entails a small loss relative to the optimum  $K_n = n$ . To our knowledge, these are the first results that provide explicit guidance on choosing  $K$  for DML under a nontrivial class of first-step estimators. Whether these conclusions extend to broader classes of first-step estimators remains an open question.

Our theoretical results support three recommendations. First, practitioners should prefer DML2 over DML1, because DML1-based inference could be invalid, and DML2 is robust to the choice of  $K$  for a large class of first-step estimators. Second, practitioners should use  $K = n$  for DML2 when the first-step estimators satisfy the conditions in Section 4 (e.g., kernel estimators). This choice of  $K$  is asymptotically optimal and it ensures replicability of the estimator. Lastly, practitioners should prefer  $K = 10$  over  $K = 5$  since this choice guarantees lower efficiency losses than  $K = 5$ .

**Related Literature:** This paper contributes to the growing DML literature, which includes 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), Ji et al. (2023), among many others. Most of the papers use DML2, with exceptions such as Chernozhukov et al. (2017), Ji et al. (2023), Noack et al. (2024), and Cheng et al. (2023), which use DML1. All these papers derive first-order fixed- $K$  asymptotic theory. In contrast, we study the asymptotic properties of DML1 and DML2 when  $K \rightarrow \infty$  as  $n \rightarrow \infty$ . We prove that DML2 offers theoretical advantages over DML1. Moreover, when the first-step estimators admitting linear stochastic expansion, we show that selecting  $K = n$  is asymptotically optimal for DML2 in terms of second-order asymptotic bias and MSE, providing theoretical guidance to choose  $K$ .

This paper also contributes to the literature on double-robust estimators, including Robins et al. (1994), Robins and Rotnitzky (1995), Scharfstein et al. (1999), Farrell (2015), Sant’Anna and Zhao (2020), Chang (2020), Callaway and Sant’Anna (2021), Rothe and Firpo (2019), and Singh and Sun (2024). All these papers study first-order asymptotic theory, with the exception of Rothe and Firpo (2019) that study higher-order properties of

double-robust leave-one-out estimators in a missing-data setting, finding that those estimators have favorable theoretical properties relative to their non-double-robust versions. We complement their work by showing that the leave-one-out estimator is optimal in terms of second-order asymptotic bias and MSE among DML2 estimators under certain conditions.

More broadly, this paper contributes to the semiparametric literature (e.g., [Bickel \(1982\)](#), [Robinson \(1988\)](#), [Newey \(1990\)](#), [Andrews \(1994\)](#), [Newey and McFadden \(1994\)](#), [Newey \(1994\)](#), [Linton \(1995\)](#), and [Bickel and Ritov \(2003\)](#)). The existing literature provides conditions to study the asymptotic properties of plug-in and leave-one-out estimators (DML2 with  $K = n$ ). We instead consider conditions for studying how to select  $K$  for DML2 through second-order asymptotic approximations.

**Outline:** Section 2 presents the setup and summarizes existing results. Section 3 presents the limiting distributions of DML1 and DML2 for large  $K$  values. Section 4 derives a second-order asymptotic approximation for DML2 allowing large  $K$  values. Section 5 presents recommendations for practitioners. Section 6 presents simulations, highlighting the relevance of our results. Section 7 concludes. Appendices A and B contain proofs of the main results and present auxiliary results. Additional proofs are in the Online Appendix.

*Notation:* We use  $[L] = \{1, \dots, L\}$ ,  $\|\cdot\|$  denotes the  $L_2$ -operator norm for matrices (and vectors),  $\|\cdot\|_\infty$  denotes the element-wise supremum norm for matrices (and vectors),  $\partial_\eta m$  denotes the matrix of partial derivatives of  $m$  with respect to  $\eta$ , and  $\mathbb{I}_d$  is the  $d \times d$  identity matrix.

## 2 Setup and Previous Results

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

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

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

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

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

where the moment function  $m$  and function  $\psi^a$  satisfy conditions specified in Assumption 3.1 in Section 3, which includes the identification condition,  $E[\psi^a(W, \eta_0(X))] \in \mathbf{R}^{d \times d}$  is

invertible, and guarantees a Neyman orthogonality condition,

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

Hereafter,  $\partial_\eta m(W, \theta_0, \eta_0(X))$  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  $\theta_0$  include the average treatment effect (Example C.1), the average treatment effect on the treated in difference-in-differences designs (Example C.2), and the local average treatment effect (Example C.3), among others. All these examples are in Appendix C and additional examples appear in Ahrens et al. (2025).

Consider the goal of estimating  $\theta_0$  using a random sample  $\{(W_i, X_i) : 1 \leq i \leq n\}$  drawn from the distribution  $F_0$ . The parameter  $\theta_0$  based on (2.1) and (2.2) can be identified as follows,

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

Accordingly, an ideal estimator for  $\theta_0$  is defined by replacing the expected values in (2.3) with sample analogs. That is,

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

where  $\eta_i = \eta_0(X_i)$  is the value of the nuisance function  $\eta_0$  for observation  $i$ . However, the values of the  $\eta_i$ 's are unknown. As a result, the ideal estimator  $\hat{\theta}_n^*$  is infeasible. For this reason, it is common to calculate first estimates  $\hat{\eta}_i$  of  $\eta_i$  that can be used later to obtain an estimator for  $\theta_0$ . Remark 2.4 discusses the *plug-in* estimator used in the semiparametric literature. In what follows, we explain how DML estimates first  $\eta_0$  and then  $\theta_0$ .

DML calculates the estimates  $\hat{\eta}_i$  of  $\eta_i$  using a *cross-fitting* procedure, which is a form of sample-splitting. This procedure has two steps and assumes that  $n$  can be divided by  $K$ :<sup>1</sup>

1. *Sample splitting*: Randomly split the indices into  $K$  equal-sized folds  $\mathcal{I}_k$ , i.e.,  $\cup_{k=1}^K \mathcal{I}_k = [n]$ . The number of observations in fold  $\mathcal{I}_k$  is denoted by  $n_k = n/K$ .
2. *Nuisance Function Estimates*: For each  $i \in \mathcal{I}_k$ , the estimates  $\hat{\eta}_i$  of  $\eta_i$  are defined by  $\hat{\eta}_i = \hat{\eta}_k(X_i)$ , where  $\hat{\eta}_k(\cdot)$  is an estimator of  $\eta_0(\cdot)$  using  $\{W_i : i \notin \mathcal{I}_k\}$ , which is all the data except the ones with indices on  $\mathcal{I}_k$ . We then repeat the process for all  $k \in [K]$ .

---

<sup>1</sup>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  $n/K$ .

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

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

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

it then combines the information across the folds by averaging the  $\tilde{\theta}_k$ 's to obtain the proposed estimator for  $\theta_0$ ,

$$\hat{\theta}_{n,K}^{(1)} = K^{-1} \sum_{k=1}^K \tilde{\theta}_k . \quad (2.5)$$

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

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

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

**Definition 2.2** (DML2). This estimator first combines the information across the folds  $\mathcal{I}_k$  by averaging the sample analog of moment conditions like (2.1) using the estimates  $\hat{\eta}_i$ , and then estimates  $\theta_0$  by solving the average of moment conditions,

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

An explicit expression for  $\hat{\theta}_{n,K}^{(2)}$  is obtained by using (2.2),

$$\hat{\theta}_{n,K}^{(2)} = \left( n^{-1} \sum_{i=1}^n \psi^a(W_i, \hat{\eta}_i) \right)^{-1} \left( n^{-1} \sum_{i=1}^n \psi^b(W_i, \hat{\eta}_i) \right) . \quad (2.6)$$

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

**Remark 2.1.** DML1 and DML2 are numerically identical when  $\psi^a(W_i, \hat{\eta}_i)$  is a constant function. This is the case for the ATE (Example C.1). In contrast, if  $\psi^a(W_i, \hat{\eta}_i)$  is not a constant, then DML1 and DML2 provide different estimates. This is the case for the ATT-DID (Example C.2) and LATE (Example C.3), among other econometric models.  $\square$

**Remark 2.2.** Chernozhukov et al. (2018, Remark 3.1) recommend using DML2 over DML1 based on simulation evidence. Specifically, they note that for some econometric models the bias and MSE of DML1 and DML2 are similar, but for others the bias and MSE of DML2 are lower than those of DML1. The literature does not provide a theoretical explanation for these findings. Section 3 will provide an explanation.  $\square$

**Remark 2.3.** Simulation evidence shows that increasing  $K$  reduces DML2's finite-sample bias and MSE (Ahrens et al., 2024a,b). An intuitive explanation is that by increasing  $K$  we improve the accuracy of first-step estimators by using more data (e.g., the first-step estimators use 50%, 80%, and 90% of the data when  $K$  is 2, 5, and 10, respectively). However, by increasing  $K$  we also increase the dependence among the first-step estimators by having more observations in common (e.g., any two first-step estimators share 0%, 60%, and 80% of the data when  $K$  is 2, 5, and 10, respectively). Then, it is unclear whether large  $K$  values provide improvement in estimation accuracy of  $\theta_0$ . Section 4 will provide an explanation when first-step estimators admit a linear stochastic expansion.  $\square$

**Remark 2.4.** Plug-in estimators for  $\theta_0$  have been considered in the semiparametric literature, but they require strong conditions to attenuate the own-observation bias. Here, by plug-in estimators, we refer to the estimator of  $\theta_0$  defined by (2.4) but with  $\eta_i$  replaced by estimates  $\hat{\eta}_i$ , where  $\hat{\eta}_i = \hat{\eta}(X_i)$  and  $\hat{\eta}$  is obtained by using all the data. The semiparametric literature has studied conditions under which it has standard properties (e.g., asymptotic normality with parametric rates), which include strong conditions on first-step estimators (e.g., a Donsker class condition) to attenuate the own-observation bias. This bias arises when the same data are used to estimate both  $\eta_0$  and  $\theta_0$  (Newey and Robins, 2018). In contrast, DML relies on general and simple conditions on first-step estimators (e.g., certain  $L_2$ -convergence rates faster than  $n^{-1/4}$ ) to obtain the standard properties. DML removes the own-observation bias by relying on cross-fitting and Neyman orthogonality.  $\square$

## 2.1 Previous Results

Under fixed- $K$  asymptotic theory, Chernozhukov et al. (2018) showed that both DML1 and DML2 have the same limiting distribution,

$$\sqrt{n} \left( \hat{\theta}_{n,K}^{(j)} - \theta_0 \right) \xrightarrow{d} N(0, \Sigma), \quad \text{for } j = 1, 2, \quad (2.7)$$

where the variance of the limiting distribution is given by

$$\Sigma = E [\psi^a(W, \eta_0(X))]^{-1} E [m(W, \theta_0, \eta_0(X))m(W, \theta_0, \eta_0(X))^{\top}] E [\psi^a(W, \eta_0(X))] , \quad (2.8)$$

which only depends on the moment function  $m$ , the true nuisance function  $\eta_0$ , and the data distribution  $F_0$ . Therefore, the existing fixed- $K$  asymptotic theory does not provide guidance for choosing between DML1 or DML2 or selecting  $K$ .

The core idea in DML is that the estimation of  $\theta_0$  using DML1 or DML2 is as accurate as if the true  $\eta_0$  had been used. Formally, both DML1 and DML2 are asymptotically equivalent to the oracle estimator  $\hat{\theta}_n^*$ ,

$$\sqrt{n} \left( \hat{\theta}_{n,K}^{(j)} - \hat{\theta}_n^* \right) \xrightarrow{p} 0, \quad j = 1, 2. \quad (2.9)$$

Although the existing asymptotic theory shows that DML1 and DML2 are asymptotically equivalent, existing simulation evidence suggests that (i) DML2 can outperform DML1, and (ii) increasing  $K$  can reduce DML2's finite-sample bias and MSE. To investigate these phenomena, we consider an asymptotic framework in which  $K$  depends on  $n$ , i.e.,  $K = K_n$ , allowing  $K_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Under standard conditions on the DML literature, we show that the bias, MSE, and inference of DML1 are sensitive to large  $K_n$  values, whereas DML2 is not, implying that DML2 offers theoretical advantages over DML1 in terms of bias, MSE, and inference. Under an additional assumption on the first-step estimators (an algorithm stability condition), we demonstrate that the limiting distribution of DML2 is not affected by the choice of  $K_n$ . Therefore, to study the effect of  $K_n$  on DML2, we further restrict the class of first-step estimators to provide a second-order asymptotic analysis. We make progress by assuming that the first-step estimators verify a linear stochastic expansion, showing that a second-order approximation for DML2 exists. We show that increasing  $K_n$  reduces the magnitude of the second-order asymptotic bias and MSE of DML2, leading to the recommendation of  $K_n = n$  for DML2 as an asymptotically optimal choice.<sup>2</sup> Lastly, our second-order asymptotic analysis finds that using  $K_n = 10$  instead of  $K_n = n$  leads to small efficiency losses in terms of second-order MSE.

### 3 Asymptotic Theory for DML1 and DML2 when $K$ increases

We show that DML2 offers theoretical advantages over DML1 (Theorem 3.1). We also prove that DML2 is robust to the choice of  $K = K_n$  for a large class of first-step estimators (Theorem 3.2). Finally, Section 3.1 explains why and when DML1 is sensitive to large  $K$ .

Let  $G = E[\psi^a(W, \eta_0(X))] \in \mathbf{R}^{d \times d}$  and  $\Omega = E[m(W, \theta_0, \eta_0(X)) m(W, \theta_0, \eta_0(X))^\top] \in \mathbf{R}^{d \times d}$ .

---

<sup>2</sup>Since  $K_n = n$  often is computationally untractable, we present in Section F.1 of the Online Appendix an informal implementation proposal for DML2 with  $K_n = n$ .

We write  $\psi^a(W, \eta) = [\psi_{t,s}^a(W, \eta)]_{t,s}$  and  $m(W, \theta, \eta) = [m_t(W, \theta, \eta)]_t$ . Recall  $\eta_i = \eta_0(X_i)$ . Let  $c_G$ ,  $c_1$ , and  $c_2$  be positive constants. The next assumption restricts the class of econometric models through conditions on the moment function  $m$  and function  $\psi^a$ .

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

- (i)  $G$  and  $\Omega$  are non-singular and  $\|G^{-1}\| < c_G$ .
- (ii)  $E[|m_t(W_i, \theta_0, \eta_i)|^4] < c_1$  and  $E[|\psi_{t,s}^a(W_i, \eta_i)|^4] < c_1$  for all  $t, s \in [d]$ .
- (iii)  $E[\partial_\eta m_t(W_i, \theta_0, \eta_i) | X_i] = 0$ ,  $\|E[(\partial_\eta m_t(W_i, \theta_0, \eta_i))(\partial_\eta m_t(W_i, \theta_0, \eta_i))^\top | X_i]\|_\infty < c_2$ , and  $\sup_{\eta \in \mathcal{E}} \|\partial_\eta^2 m_t(W_i, \theta_0, \eta)\|_\infty \leq c_2$  for  $t \in [d]$ .
- (iv)  $E[\partial_\eta \psi_{t,s}^a(W_i, \eta_i) | X_i] = 0$ ,  $\|E[(\partial_\eta \psi_{t,s}^a(W_i, \eta_i))(\partial_\eta \psi_{t,s}^a(W_i, \eta_i))^\top | X_i]\| < c_2$ , and  $\sup_{\eta \in \mathcal{E}} \|\partial_\eta^2 \psi_{t,s}^a(W_i, \eta)\|_\infty \leq c_2$  for all  $t, s \in [d]$ .

Parts (i) and (ii) of Assumption 3.1 are standard conditions that guarantee identification of the parameter of interest and stochastic expansions for the oracle estimator, similar to Newey and Smith (2004, Assumptions 2 and 3). Part (iii) presents a Neyman-orthogonality condition,  $E[\partial_\eta m_t(W_i, \theta_0, \eta_i) | X_i] = 0$ , involving standard partial derivatives as in Belloni et al. (2017) and Farrell et al. (2025) rather than functional derivatives as in Chernozhukov et al. (2018). In general, these type of conditions are necessary to guarantee that feasible estimators are as accurate as if the true values of the nuisance functions were used; see Remark 3.1 for an explanation. It is possible to transform a moment function into one that satisfies a Neyman-orthogonality condition under certain conditions; see Remark 3.2 for comments on existing methods. Part (iii) also includes standard conditions ensuring that nonlinear effects of first-step estimation error are negligible when they are sufficiently accurate (e.g., when their  $L_2$ -convergence rates are faster than  $n^{-1/4}$ ). Finally, part (iv) of Assumption 3.1 implies that we can construct a DML estimator for each component of  $G = E[\psi^a(W_i, \eta_i)]$ . It holds automatically when  $\psi^a$  does not depend on  $\eta$ . We use part (iv) only for the analysis of DML1 and not for DML2.

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

Let  $\tau_n$  be a sequence of positive numbers converging to zero. Let  $n_0 = (1 - 1/K_n)n$  be the number of observations in the sample  $\{W_i : i \notin \mathcal{I}_k\}$  used by  $\hat{\eta}_k$  to estimate  $\eta_0$ . We next present conditions on the class of first-step estimators.

**Assumption 3.2** ( $L_2$ -convergence rate).  $E [||\hat{\eta}_k(X) - \eta_0(X)||^2]^{1/2} \leq n_0^{-1/4} \tau_{n_0}$ .

Assumption 3.2 holds for several first-step estimators considered in the DML literature. For instance, it holds for deep neural networks as in Farrell et al. (2021) and Schmidt-Hieber (2020). Under certain conditions, it holds for LASSO and related penalized estimators of linear models (Tibshirani (1996), Van de Geer (2008), Belloni et al. (2011)). Kernel estimators and series estimators (Newey (1997), Belloni et al. (2015), Chen (2007)) can verify this assumption after an appropriate trimming to ensure bounded inverse density weights for kernel estimators, or well-conditioned Gram matrices for series estimators. It also holds for an honest version of random forest (Chernozhukov et al., 2024, Theorem 9.4.3); see Chi et al. (2022) for convergence rates of high-dimensional random forest.

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

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

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

and

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

where  $\hat{\theta}_{n, K_n}^{(1)}$ ,  $\hat{\theta}_{n, K_n}^{(2)}$ , and  $\Sigma$  are defined in (2.5), (2.6), and (2.8), respectively, and

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

Theorem 3.1 provides an asymptotic result that explains the discrepancy found in simulations between DML1 and DML2. As we note in Remark 2.2, DML2 has been recommended over DML1 based on simulation evidence. This theorem now provides the theoretical explanation. It shows that DML2 is asymptotically better than DML1 in terms of bias and MSE when  $c > 0$  and  $\Lambda \neq 0$ ; otherwise, both share the same limiting distribution. Our explanation through  $\Lambda$  emerges under the proposed asymptotic framework, providing insights not captured by the existing fixed- $K$  asymptotic theory or simulation-based evidence.

The distinction between DML1 and DML2 relies only on the model-dependent quantity  $\Lambda$  and not on the first-step estimator  $\hat{\eta}$ . Therefore, the relative performance between DML1 and DML2 can be obtained by calculating  $\Lambda$  without running simulations. In particular,

for several econometric models—such as ATE, ATT-DID, and PLM— $\Lambda = 0$ , but for others like LATE and PLM-IV, it is typically nonzero. Note that  $\Lambda$  can be calculated also for econometric models where the moment function is not differentiable in  $\eta$ . We argue that if for those models  $\Lambda \neq 0$ , then DML2 should be preferred over DML1 even if those models are outside the scope of our setup. We postpone our explanation to Section 3.1.

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

**Corollary 3.1.** *Let  $CI_\alpha^{(1)}$  be the standard recommended DML1 confidence interval for  $\theta_{0,t}$  ( $t$ -th component of  $\theta_0$ ),*

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

where  $\hat{s}_{t,n}^2$  is a consistent estimator for  $\Sigma_{t,t}$  ( $t$ -th diagonal term of  $\Sigma$ ). Under the conditions of Theorem 3.1, we then have  $P\left(\theta_{0,t} \in CI_\alpha^{(1)}\right) = \mathcal{P}(K_n) + o(1)$ , where  $\mathcal{P}(K_n) = \Phi\left(z_{1-\alpha/2} + \frac{K_n}{\sqrt{n}} \frac{\Lambda_t}{\sqrt{\Sigma_{t,t}}}\right) + \Phi\left(z_{1-\alpha/2} - \frac{K_n}{\sqrt{n}} \frac{\Lambda_t}{\sqrt{\Sigma_{t,t}}}\right) - 1$ . Here,  $\Lambda = [\Lambda_t]_t$  and  $\Sigma = [\Sigma_{t,s}]_{t,s}$ . Furthermore,  $\mathcal{P}(K_n)$  is a decreasing function on  $K_n$  if and only if  $\Lambda_t \neq 0$ . In particular,  $\mathcal{P}(0) = 1 - \alpha > \mathcal{P}(K_n)$  if  $K_n > 1$  and  $\Lambda_t \neq 0$ .

Theorem 3.1 has shown that DML2 asymptotically dominates DML1 and is robust to the choice of  $K_n$ , provided that  $K_n = O(\sqrt{n})$ . The next assumption is proposed to extend the robustness of DML2 to the choice of  $K_n$  from  $K_n = O(\sqrt{n})$  to  $K_n = O(n)$ . Let  $\hat{\eta}_k^\ell(\cdot)$  be the same estimator as  $\hat{\eta}_k(\cdot)$  except in the use of observation  $\ell$ :  $\hat{\eta}_k(\cdot)$  uses  $(W_\ell, X_\ell)$ , whereas  $\hat{\eta}_k^\ell(\cdot)$  uses  $(\widetilde{W}_\ell, \widetilde{X}_\ell)$ , where the random vector  $(\widetilde{W}_\ell, \widetilde{X}_\ell)$  is drawn from  $F_0$  and independent of the data (i.e.,  $(\widetilde{W}_\ell, \widetilde{X}_\ell)$  and  $(W_\ell, X_\ell)$  are i.i.d.). Recall  $\tau_{n_0} = o(1)$  and  $n_0 = (1 - 1/K_n)n$ .

**Assumption 3.3** (Algorithm stability).  $\max_{\ell \notin \mathcal{I}_k} E[|\hat{\eta}_k(X) - \hat{\eta}_k^\ell(X)|^2]^{1/2} \leq n_0^{-1/2} \tau_{n_0}$

Assumption 3.3 measures the stability of first-step estimators to replacement of exactly one observation in  $L_2$ -norm. It is a sufficient condition that we use to control the dependence among the prediction errors of the moment function due to first-step errors,  $\{m(W_i, \theta_0, \hat{\eta}_i) -$

$m(W_i, \theta_0, \eta_i) : 1 \leq i \leq n$ . A relatively stronger version of this assumption appears in [Chen et al. \(2022, Corollary 4\)](#). Kernel and series estimators verify this condition after appropriate trimming. LASSO can also verify this condition under an strong sparsity condition. We do not know whether this condition holds for deep neural networks or other machine-learning methods; characterizing when it holds remains an open question.

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

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

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

where  $\hat{\theta}_{n, K_n}^{(2)}$  and  $\Sigma$  are as in (2.6) and (2.8), respectively.

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

One of the main benefits of using DML2 with  $K_n = n$  is that it ensures replicability. DML2 with  $K_n = n$  is uniquely determined by the data and therefore eliminates the random-split variability that exists when  $K_n < n$ , where different random splits yield different DML2 estimates. However, its implementation in practice may appear challenging due to the computational burden of estimating  $n$  first-step estimators. Section F.1 in the Online Appendix proposes an informal implementation for DML2 with  $K_n = n$ .

The next corollary formalize that DML2-based inference is robust to the choice of  $K_n$ .

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

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

where  $CI_{\alpha}^{(2)}$  is the standard recommended DML2 confidence interval for  $\theta_{0,t}$  ( $t$ -th component of  $\theta_0$ ),

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

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

An important caveat is that our results so far do not provide guidance on the choice of  $K_n$ . Our first-order asymptotic theory shows that, under our assumptions,  $K_n$  does not affect the approximation of the finite-sample distribution of DML2. Therefore, in Section 4, we make progress on the selection of  $K_n$  by using stronger conditions on the class of first-step estimators. Specifically, for first-step estimators that admit linear stochastic expansions, we derive a second-order asymptotic approximation to explain the finite-sample bias and MSE of DML2, following a long tradition in econometrics of using second-order asymptotic approximations to compare first-order asymptotically equivalent estimators (Rothenberg, 1984; Linton, 1995; Donald and Newey, 2001; Newey and Smith, 2004).

**Remark 3.1.** A Neyman orthogonality condition of the moment function is necessary to guarantee a first-order equivalent condition between a feasible estimator and its oracle version (as in (2.9)). To see this, consider the following example. Suppose  $\psi^a(W, \eta) = 1$  and  $\psi^b(W, \eta)$  is a linear function in  $\eta$ . In addition, assume that the nuisance parameter  $\eta_0$  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  $\eta_0$  such that  $n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{d} N(0, V_\eta)$  and  $V_\eta$  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 part (c) of Assumption 3.1 holds. See also Andrews (1994, Eq. (2.12)).  $\square$

**Remark 3.2.** We can obtain moment functions satisfying Neyman orthogonality by adding adjustment terms to the original moment functions. The adjustment terms are constructed using first-order influence functions, as developed in Newey (1994), Hahn and Ridder (2013), Ichimura and Newey (2022), and Farrell et al. (2025), among others. Since the analytical construction can be tedious, the DML literature has focused on automatic construction of orthogonal moments: procedures that take the original moment function as input and automatically return a DML2 estimate. Examples include Chernozhukov et al. (2022a), Escanciano and Pérez-Izquierdo (2023), and Argañaraz (2025).  $\square$

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

### 3.1 Why and When DML1 is Sensitive to $K$ Increasing

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

The oracle version of DML1 is defined as DML1 but using  $\eta_i$ 's instead of  $\hat{\eta}_i$ , that is, assuming perfect knowledge of  $\eta_0$ ,

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

where  $\tilde{\theta}_k^*$ 's are preliminary estimators solving the moment condition (2.1) within each fold  $\mathcal{I}_k$  and using the true values of  $\eta_i$ ,

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

Standard arguments (Newey and Smith, 2004) show that the second-order asymptotic bias of  $\tilde{\theta}_k^*$  is  $\Lambda K n^{-1/2}$ , which implies that the second-order asymptotic bias of  $\hat{\theta}_{n,K}^{*,(1)}$  is also  $\Lambda K n^{-1/2}$ , since the oracle DML1 is an average of those preliminary estimators. When  $K \propto n^{1/2}$  and  $\Lambda \neq 0$ , the second-order asymptotic bias becomes first-order and prevents correct centering of the limiting distribution of the oracle DML1. We formalize this next.

**Theorem 3.3.** *Let Assumption 3.1 (i)–(ii) hold and let  $K_n$  be such that  $K_n \leq n$  and  $K_n/\sqrt{n} \rightarrow c \in [0, \infty)$  as  $n \rightarrow \infty$ . Then,*

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

where  $\Sigma$  and  $\Lambda$  are defined in (2.8) and (3.1), respectively.

This theorem holds for a large class of econometric models, including non-smooth ones. We use mild regularity conditions to derive the limiting distribution of the oracle DML1. In particular, we don't rely on the smoothness of the moment function  $m$  with respect to  $\eta$ .

Theorem 3.3 shows that the oracle DML1 has the same limiting distribution that we derive for DML1 in Theorem 3.1. This last result occurs because the proof of Theorem 3.1 uses that the DML1 and the oracle DML1 are asymptotically equivalent,

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

Part (d) of Assumption 3.1 is key to guarantee that (3.3) holds.

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

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

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

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

The asymptotic equivalence between DML2 and its oracle version is sufficient to conclude the robustness of DML2 to the choice of  $K$ ,

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

We show in part (ii) of Lemma A.1 that Assumptions 3.1 (i)–(iii) and 3.2 are sufficient to verify the asymptotic equivalence in (3.5), provided that  $K_n = O(\sqrt{n})$ . To guarantee that (3.5) continues to hold for  $K_n \propto n$  we additionally use Assumption 3.3, we formalize this in Lemma A.2. We do not require Assumption 3.1 (iv) for the analysis of DML2.

Finally, notice that the oracle version of DML1 depends on random splitting, while the oracle version of DML2 does not. Therefore, even if we use the oracle DML1, it lacks replicability due to sample splitting, whereas the oracle DML2 does not.

## 4 Second-Order Asymptotic Approximation for DML2 when $K$ increases

Section 3 shows that the limiting distribution of DML2 is invariant to  $K_n$ , so first-order asymptotics provide no guidance on the choice of  $K_n$ , thereby motivating a second-order asymptotic approximation to understand how  $K_n$  affects DML2. To derive such an approximation (Theorem 4.1) and obtain guidance on choosing  $K_n$ , we focus on scalar DML2 estimators and first-step estimators that admit a linear stochastic expansion (e.g., kernel estimators with MSE-optimal bandwidths). Using the second-order asymptotic bias and MSE in Theorem 4.2, we characterize the optimal choice of  $K_n$  and quantify the relative efficiency loss from suboptimal choices. Under our conditions, the second-order asymptotic

bias and MSE of DML2 decrease with  $K_n$ , implying that  $K_n = n$  is an asymptotically optimal choice within this setting. To our knowledge, these are the first results that provide explicit guidance on choosing  $K_n$  for DML under a nontrivial class of first-step estimators.

Let  $n_0^{-\varphi}$  be the  $L_2$ -convergence rate of the estimator  $\hat{\eta}_k$ , where  $n_0 = (1 - 1/K_n)n$  is the number of observations in the sample  $\{W_i : i \notin \mathcal{I}_k\}$  used by  $\hat{\eta}_k$  to estimate  $\eta_0$ . Let  $M_1$ ,  $M_2$ ,  $c_\delta$ , and  $c_b$  be positive constants. Recall  $\tau_n = o(1)$ . To derive the second-order asymptotic approximation for DML2, the next assumption requires a linear stochastic expansion for the first-step estimator, which holds for kernel estimators with MSE-optimal bandwidths. It is unknown whether LASSO or deep neural networks have stochastic expansions.

**Assumption 4.1** (Linear stochastic expansion). *There exist  $\delta_{n_0} : \mathcal{W} \times \mathcal{X} \rightarrow \mathbf{R}^p$  and  $b_{n_0} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}^p$ , such that*

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

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

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

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

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

Part (i) of Assumption 4.1 is a high-level condition that presents a linear stochastic expansion for the first-step estimator  $\hat{\eta}_k$ , with variance and bias contributions given by  $\delta_{n_0}$  and  $b_{n_0}$ , respectively. Part (ii) is a standard requirement in the semiparametric literature (Bickel and Ritov, 2003). Parts (iii) and (iv) impose regularity conditions on  $\delta_{n_0}$  and  $b_{n_0}$  that imply  $n_0^{-2\varphi}$  is the convergence rate of both the squared bias and variance of  $\hat{\eta}_k$ .

Assumption 4.1 further restricts the class of first-step estimators considered in Section 3, in the sense that we can verify that this assumption implies both Assumptions 3.2 and 3.3. Furthermore, this assumption provides additional structure on the estimators  $\hat{\eta}_k$  that we can use to derive a second-order asymptotic approximation for the scalar DML2 estimator. Nevertheless, conducting an appropriate analysis of the leading terms of the second-order bias and MSE of DML2 requires additional conditions on the functions  $\delta_{n_0}$  and  $b_{n_0}$  and the higher-order partial derivatives of the moment function  $m$ . We formalize those conditions in the next assumption. To simplify notation, let  $\tilde{b}_{n_0}(X_i) = E[b_{n_0}(X_j, X_i) | X_i]$  for  $j \neq i$ , let  $\partial_\eta^2 m_i = \partial_\eta^2 m(W_i, \theta_0, \eta_i)$ , and recall that  $\eta_i = \eta_0(X_i)$  and  $G = E[\psi^a(W_i, \eta_i)]$ .

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

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

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

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

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

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

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

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

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

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

Theorem 4.1 demonstrates that  $\mathcal{T}_{n,K_n}$  provides a better asymptotic approximation than  $\mathcal{T}_n^*$ . We obtain this improvement by including  $\mathcal{T}_{n,K_n}^{nl}$  to account for nonlinear effects of first-step estimation errors. The proof of this theorem uses that  $\mathcal{T}_{n,K_n}^{nl}$  is the leading term in the scaled difference between  $\hat{\theta}_{n,K_n}^{(2)}$  and its oracle version  $\hat{\theta}_{n,K_n}^{*,(2)}$  defined in (3.4):

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

**Remark 4.1.** When  $K_n$  is fixed as  $n \rightarrow \infty$ , the second-order asymptotic approximation for DML1 is also  $\mathcal{T}_{n,K_n}$ , which is identical to the one obtained for DML2 in (4.5), making it impossible to distinguish them using second-order asymptotic analysis and fixed  $K_n$ . Thus, distinguishing DML1 from DML2 requires an asymptotic framework where  $K_n \rightarrow \infty$  as  $n \rightarrow \infty$ , as we develop in Section 3.  $\square$

**Remark 4.2.** The second-order asymptotic approximation  $\mathcal{T}_{n,K_n}$  for DML2 depends on the sample-splitting partition through  $\mathcal{T}_{n,K_n}^{nl}$ ; in contrast, the first-order asymptotic approximation  $\mathcal{T}_n^*$  does not. Therefore,  $\mathcal{T}_{n,K_n}$  can be used to study the effects of sample splitting in DML2, which is an interesting research direction left for future work.  $\square$

## 4.1 Second-order Asymptotic bias and MSE for DML2

We next use the asymptotic approximation  $\mathcal{T}_{n,K_n}$  to explain how the choice of  $K_n$  affects the bias and MSE of DML2. We define the second-order asymptotic bias and MSE of DML2 as the mean and second moment of  $\mathcal{T}_{n,K_n}$ , respectively. These definitions follow a long tradition in econometrics of using second-order approximations to compare estimators with identical first-order asymptotic properties (Rothenberg, 1984; Linton, 1995; Newey and Smith, 2004). Recall that  $\varphi \in (1/4, 1/2)$  by Assumption 4.1.

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

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

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

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

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

Theorem 4.2 provides an asymptotic result that explains observed patterns in DML2's finite-sample bias and MSE when first-step estimators satisfy our conditions (Remark 2.3). Since  $\mathcal{B} \neq 0$  and  $\mathcal{C} > 0$ , we see that the magnitude of the second-order asymptotic bias and MSE decrease with  $K_n$ , consistent with existing simulations findings (Ahrens et al., 2024a,b), and our simulation results in Section 6.2. The simulation results in Section 6.2 show that

for  $K_n \geq 10$ , DML2's finite-sample bias and MSE appear approximately constant, which is consistent with the fact that the terms  $(1 + 1/(K_n - 1))^{2\varphi}$  and  $(1 + 1/(K_n - 1))^{2\varphi-1/2}$  in (4.6)–(4.7) change little when  $K_n \geq 10$  for typical values  $\varphi \in (1/4, 1/2)$ .

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

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

for any sequence  $K_n$  such that  $K_n \leq n$  since  $\mathcal{C} > 0$ .

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

The previous result demonstrates that the leave-one-out estimator, which is DML2 with  $K_n = n$ , is optimal among DML2's in terms of second-order asymptotic MSE. We can perform a similar analysis using the magnitude of the second-order asymptotic bias of DML2 as our optimality criterion, with analogous results, concluding that  $K_n = n$  is also an optimal choice as long as  $\mathcal{B} \neq 0$ . Therefore, the leave-one-out estimator is also optimal among DML2's in terms of second-order asymptotic bias.

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

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

where  $\Sigma$  and  $\Lambda$  are defined in (2.8) and (3.1), respectively, and

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

□

## 4.2 Relative efficiency loss from suboptimal choice of $K$

In the remainder of this section, we quantify the relative efficiency loss from any suboptimal choice of  $K$  using Theorem 4.2. The main motivation is to evaluate the performance of commonly recommended choices such as  $K = 5$  or  $K = 10$  relative to the optimal choice.

The relative efficiency loss of the choice  $K$  in terms of second-order asymptotic MSE is

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

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

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

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

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

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

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

The relative efficiency loss of the choice  $K$  in terms of second-order asymptotic bias is

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

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

From (4.9), we conclude that  $\mathcal{RL}_{\text{bias}}(5) \in (11.7\%, 24.9\%)$  and  $\mathcal{RL}_{\text{bias}}(10) \in (5.4\%, 11\%)$  when  $\varphi \in (1/4, 1/2)$  and  $n \geq 1000$ ; therefore, the relative efficiency loss from commonly recommended choices such as  $K = 5$  or  $10$  may be significant in terms of second-order asymptotic bias. Furthermore, these relative losses increase as the first-step estimator be-

comes more accurate (i.e., as  $\varphi$  increases), showing that optimal choice of  $K$  in terms of bias is critical when nuisance estimation is faster.

## 5 Recommendations for Practitioners

Our recommendations focus on two key decisions: DML1 versus DML2, and the choice of  $K$ . To support the first decision, we use the results in Section 3. For the choice of  $K$ , we rely on the theory in Section 4, which considers a restricted class of first-step estimators. Whether our recommendation for  $K$  extends to a broader class of first-step estimators—such as LASSO or deep neural networks—remains an open question. In ongoing work, Monte Carlo simulations suggest that our recommendations for  $K$  apply more broadly.

### 5.1 Prefer DML2 over DML1

Practitioners should use DML2 over DML1. While this is consistent with existing practice, we offer several supporting reasons. First, DML2 offers theoretical advantages over DML1 in terms of bias and MSE (Theorem 3.1). Second, DML1-based inference is invalid when  $\Lambda \neq 0$  (Corollary 3.1), where  $\Lambda$  is a model-dependent quantity defined in (3.1). Third, DML2 is robust to the choice of  $K$  for a large class of first-step estimators (Theorem 3.2). Fourth, estimation and inference using DML2 are reliable for any choice of  $K \in \{2, \dots, n\}$  (Corollary 3.2). Lastly, DML2 is simpler to use than DML1; DML2 doesn't require verifying if  $\Lambda$  equals zero or not, and it has available implementations for many econometric settings in *Stata* (*ddml*; Ahrens et al. (2024a)), Python (*DoubleML*; Bach et al. (2022)), and R (*DoubleML*; Bach et al. (2024)).

### 5.2 Use $K = n$ for DML2 with classical nonparametric first-steps

Practitioners should use DML2 with  $K = n$  when the first-step estimators satisfy the conditions in Section 4 (e.g., kernel estimators). Two reasons supporting this choice. First,  $K = n$  is an asymptotically optimal choice for DML2 in terms of second-order asymptotic bias and MSE. Second,  $K = n$  ensures replicability by eliminating random-split variability. In contrast, for any  $K < n$ , different random splits yield different DML2 estimates, so researchers analyzing the same data with the same  $K$  could obtain different conclusions.

### 5.3 Use $K = 10$ over $K = 5$

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

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

## 6 Simulations

This section examines how well the asymptotic approximations in Sections 3 and 4 capture finite-sample behavior in two models: (i) ATT-DID (Sant’Anna and Zhao, 2020) and (ii) LATE (Hong and Nekipelov, 2010). We calculate the bias, MSE, and coverage probability of confidence intervals for DML1 and DML2 for several values of  $K$ . We use confidence intervals based on Chernozhukov et al. (2018, Theorem 3.2).

### 6.1 Design: ATT-DID and LATE

**ATT-DID:** This section is based on Example C.2. We build 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

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

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

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

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.,  $X_{1,i} \sim \text{Uniform}[0, 1]$ ).

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

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

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

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

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  $\xi_{4,i} \sim \text{Poisson}(1)$ . All of them are independent conditional on  $X_i$ . The treatment assignment is defined by  $Z_i \sim \text{Bernoulli}(\Phi(X_i - 0.5))$ .

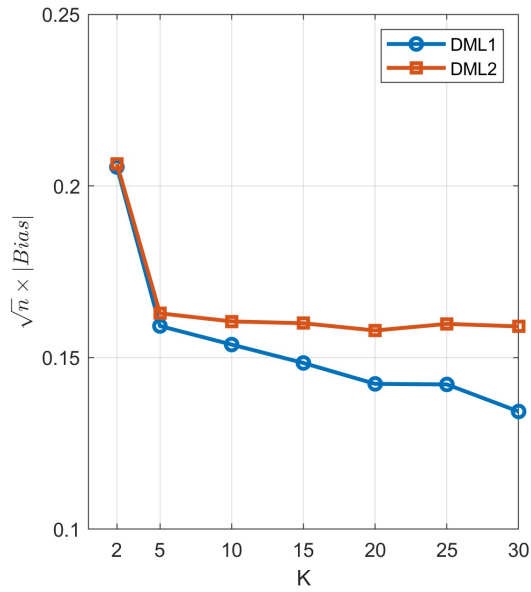
## 6.2 Results: LATE is sensitive to large $K$ values, while ATT-DID is not

This section provides simulation evidence showing that DML2 strictly dominates DML1 in the case of LATE, but performs similarly for the case of ATT-DID. This is consistent with our results in Section 3 since LATE has  $\Lambda \neq 0$ , while ATT-DID has  $\Lambda = 0$ .

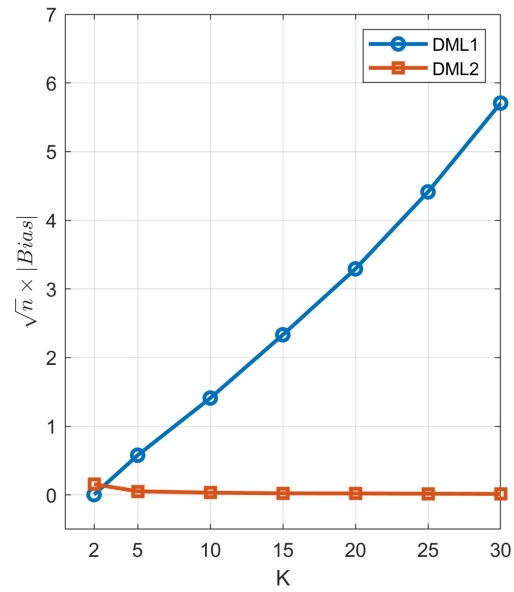
We calculate DML1 and DML2 for the ATT-DID (Example C.2) and LATE (Example C.3) for different values of  $K \in \{2, 5, 10, 15, 20, 25, 30\}$ . The nuisance function  $\eta_0$  for the ATT-DID and LATE are presented in Examples C.2 and C.3, respectively. We estimate each component of  $\eta_0$  using Nadaraya-Watson estimators and cross-fitting (see Section 2), where each first-step estimator uses sample size  $n_0 = ((K - 1)/K)n$ . For the ATT-DID, we use a 6th-order Gaussian kernel and common bandwidth  $h_j = cn_0^{-1/16}$  for all coordinates.<sup>3</sup> For the LATE, we use a 2nd-order Gaussian kernel and bandwidth  $h_j = cn_0^{-1/5}$ .

---

<sup>3</sup>We also considered a 2nd order Gaussian Kernel in the simulations. The results are presented in Figures F.1 and F.2 in Online Appendix F, and they are similar to the ones presented using a 6th order Gaussian kernel.

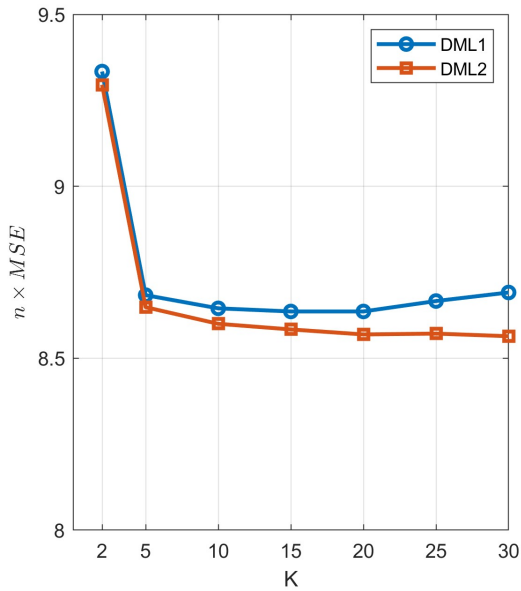


(a) ATT-DID ( $\Lambda = 0$ )

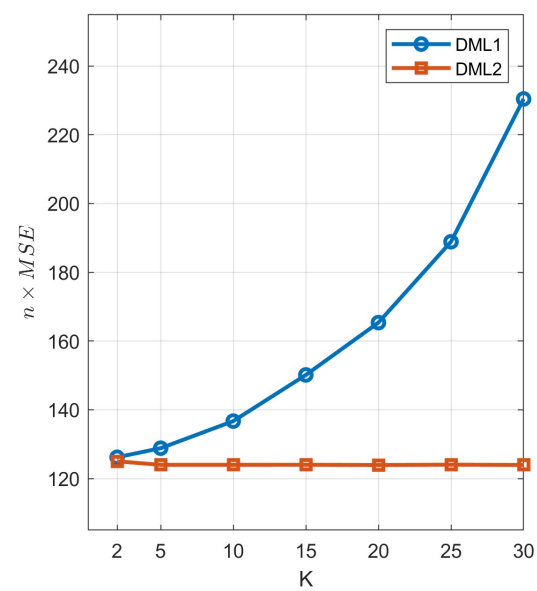


(b) LATE ( $\Lambda \neq 0$ )

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



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



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

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

**Bias:** Figure 1 presents the bias of DML1 and DML2 for several values of  $K$  and two models: ATT-DID in panel (a) and LATE in panel (b). Panel (a) shows that DML1 and DML2 perform similarly in terms of bias, while panel (b) shows that the bias of DML1 grows almost linearly in  $K$ . Theorem 3.1 explains this finite-sample behavior since  $\Lambda = 0$  for panel (a), while  $\Lambda \neq 0$  for panel (b). Note that in both panels, the bias of DML2 decreases in  $K$  and remains approximately constant for  $K \geq 10$ . We can explain this phenomenon by noting that  $(1 + 1/(K - 1))^{2\varphi}$  in (4.6) changes little when  $K \geq 10$ .

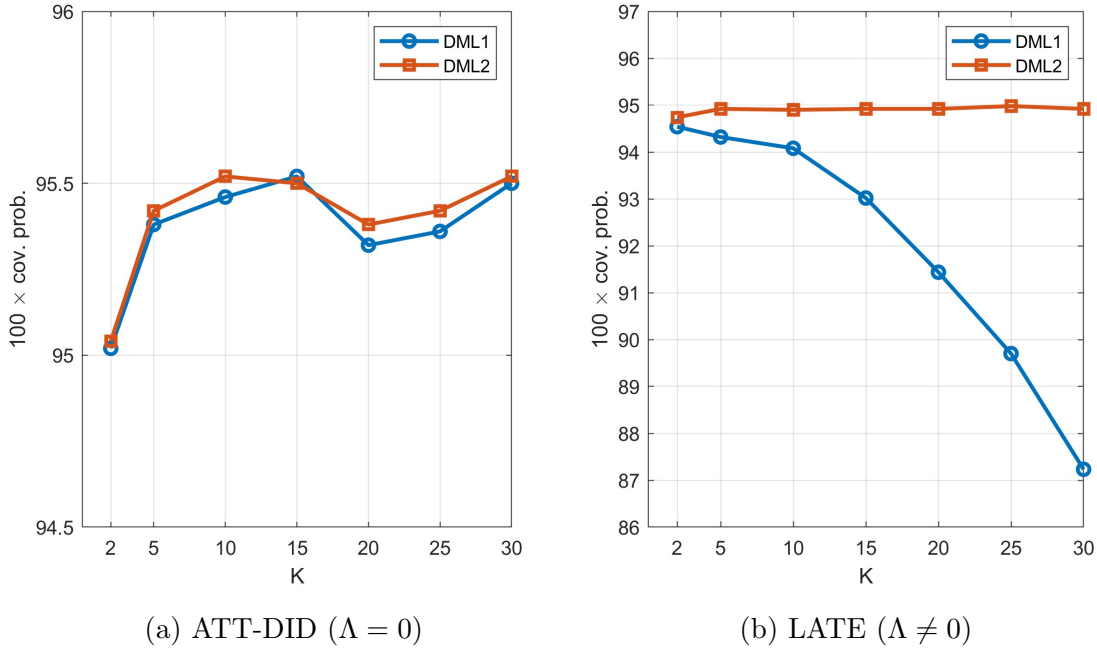


Figure 3: Coverage probability of 95%-confidence intervals based on DML1 and DML2 for ATT-DID and LATE. Sample size  $n = 3,000$  and 5,000 simulations.

**MSE:** Similarly, Figure 2 presents the MSE of DML1 and DML2 across several values of  $K$  for two models. Panel (a) shows that DML1 and DML2 perform similarly in terms of MSE, consistent with Theorem 3.1 since  $\Lambda = 0$  in this case. In contrast, panel (b) corresponds to a case where  $\Lambda \neq 0$ . It shows that the MSE of DML1 increases approximately quadratically in  $K$ . This finding aligns with the expressions in Remark 4.3 for the oracle version of DML1. Additional simulation results in Figure F.3 in the Online Appendix F show that DML1 and the oracle DML1 exhibit similar MSE values. Note that in both panels, the MSE of DML2 decreases in  $K$  and remains approximately constant for  $K \geq 10$ . We can explain this phenomenon by noting that  $(1 + 1/(K - 1))^{2\varphi-1/2}$  in (4.7) changes little when  $K \geq 10$ .

**Coverage probability:** Figure 3 presents coverage probability results for 95% confidence intervals based on DML1 and DML2 across several values of  $K$  for two models: ATT-DID in panel (a) and LATE in panel (b). Panel (a) corresponds to a case where  $\Lambda = 0$  and shows similar coverage for DML1 and DML2, whereas panel (b) corresponds to a case where  $\Lambda \neq 0$  and shows DML1 coverage distortions increasing with  $K$ . Corollaries 3.1 and 3.2 explain the finite-sample behavior observed in both panels.

**Remark 6.1.** Figures F.5 and F.6 in the Online Appendix F show that the bias and MSE could be non-monotonic on  $K$  for certain bandwidths. Concretely, we report additional results for the ATT-DID and LATE for different choices of the bandwidths.  $\square$

## 7 Concluding remarks

This paper studies DML1 and DML2 under a novel asymptotic framework in which the number of folds  $K$  can grow with the sample size  $n$ . Under this framework, we explain why and when DML2 outperforms DML1, thereby formalizing existing simulation evidence. We show that a model-based quantity  $\Lambda$  characterizes their first-order asymptotic difference: when  $\Lambda = 0$ , DML1 and DML2 perform comparably, whereas when  $\Lambda \neq 0$ , DML1 becomes sensitive to large values of  $K$ , but DML2 does not. This insight is not captured by existing fixed- $K$  asymptotic theory or simulation-based evidence alone.

Beyond explaining these patterns, our results provide guidance for practitioners, with some caveats on the choice of  $K$ . For the decision between DML1 vs DML2, we offer several reasons to prefer DML2 over DML1, formalizing the existing recommendation in the literature. Among them: inference based on DML1 can be invalid for large  $K$  values, whereas inference based on DML2 is robust to the choice of  $K$  for a large class of first-step estimators. The robustness of DML2 to choices of  $K \propto n$  relies on an algorithm stability condition, which is a condition requiring that first-step estimators are stable to replacing a single observation with another i.i.d. draw. To our knowledge, it is unknown whether this condition holds for deep neural networks or other machine-learning methods; characterizing when it holds remains an open question.

We make progress on the choice of  $K$  by restricting attention to first-step estimators that admit a linear stochastic expansion. Our second-order asymptotic analysis shows that, under the conditions we assume, the second-order bias and MSE of DML2 decrease with  $K$ , making  $K = n$  an asymptotically optimal choice for DML2 and also fully replicable. For practitioners who cannot implement  $K = n$ , we provide a theoretical basis for preferring  $K = 10$  over  $K = 5$ , with an efficiency loss in terms of second-order MSE of at most

5.4% rather than 11.8%. Whether our recommendations for  $K$  extend to a broader class of first-step estimators is left for future work.

## A Proof of Main Results

### A.1 Proofs for Section 3

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

(i) *Then, equation (3.3) holds.*

(ii) *Then, equation (3.5) holds.*

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

*Proof of Theorem 3.1.* First, Lemma A.1 implies that, for  $j = 1, 2$ , we have

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

Second, since  $\hat{\theta}_{n,K_n}^{*,(2)} = \hat{\theta}_n^*$ , we conclude  $\sqrt{n}(\hat{\theta}_{n,K_n}^{*,(2)} - \theta_0) \xrightarrow{d} N(0, \Sigma)$  by standard arguments. Finally, Theorem 3.3 in Section 3.1 demonstrates that  $\sqrt{n}(\hat{\theta}_{n,K_n}^{*,(1)} - \theta_0) \xrightarrow{d} N(c\Lambda, \Sigma)$ .  $\square$

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

*Proof of Theorem 3.3.* We use the definition of  $\hat{\theta}_{n,K_n}^{*,(1)}$  to write

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

where  $a_k$  and  $b_k$  are defined in Appendix B.

We next apply identity (B.1) from Appendix B twice to write

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

where

$$I_1 = K_n^{-1/2} \sum_{k=1}^{K_n} a_k$$

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

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

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

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

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

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

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

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

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

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

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

### A.1.1 Proof of Lemma 1

Part (i): Let  $\hat{a}_k$ ,  $\hat{b}_k$ ,  $a_k$ , and  $b_k$  denote the quantities defined in Appendix B. We write  $\sqrt{n} \left( \hat{\theta}_{n, K_n}^{(1)} - \hat{\theta}_{n, K_n}^{*, (1)} \right) = A + B$ , where

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

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

The identity (B.1) and triangle inequality imply

$$\|A\| \leq \|I_1\| + \max_{1 \leq k \leq K_n} \left\| \left( \mathbb{I}_d + n_k^{-1/2} \hat{b}_k \right)^{-1} \right\| \times (I_2 + I_3)$$

where  $I_2 = n^{-1/2} \sum_{k=1}^{K_n} \left\| \hat{b}_k - b_k \right\| \times \|\hat{a}_k - a_k\|$ ,  $I_3 = n^{-1/2} \sum_{k=1}^{K_n} \|b_k\| \times \|\hat{a}_k - a_k\|$ , and

$$I_1 = K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k . \tag{A.1}$$

The inequality (B.2) and triangle inequality imply

$$\|B\| \leq \max_{1 \leq k \leq K_n} \left\| \left( \mathbb{I}_d + n_k^{-1/2} b_k \right)^{-1} \right\| \times \max_{1 \leq k \leq K_n} \left\| \left( \mathbb{I}_d + n_k^{-1/2} \hat{b}_k \right)^{-1} \right\| \times I_4 ,$$

where  $I_4 = n^{-1/2} \sum_{k=1}^{K_n} \left\| \hat{b}_k - b_k \right\| \times \|a_k\|$ .

We next show that  $I_j = o_p(1)$  for  $j = 1, 2, 3, 4$ , which is sufficient to complete the proof of part (i) since Lemma B.2 guarantees that both  $\max_{1 \leq k \leq K_n} \left\| \left( \mathbb{I}_d + n_k^{-1/2} b_k \right)^{-1} \right\|$  and  $\max_{1 \leq k \leq K_n} \left\| \left( \mathbb{I}_d + n_k^{-1/2} \hat{b}_k \right)^{-1} \right\|$  are  $O_p(1)$  when  $K_n = O(n^{1/2})$ .

**Claim 1:**  $I_1 = o_p(1)$ . We use Taylor expansion with Lagrange remainder term for each coordinate and notation defined in Appendix B to write  $I_1 = I_{1,1} + I_{1,2}$ , where

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

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

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

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

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

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

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

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

**Claim 2:**  $I_2 = o_p(1)$ . We first use Taylor expansion with Lagrange remainder term for each coordinate and notation defined in Appendix B to write

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

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

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

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

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

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

where (1) holds by the triangle inequality and Lemma B.1, (2) holds by Lemma B.2 and the Cauchy-Schwarz inequality, (3) holds by part (ii) of Assumption 3.1, using the definition of  $b_k$ , and since  $K_n = O(n^{1/2})$ , and (4) holds by Lemma B.3 and since  $K_n = O(n^{1/2})$ .

**Claim 4:**  $I_4 = o_p(1)$ . The proof is similar to Claim 3 but using  $\mathcal{D}\psi_n^a$  instead of  $\mathcal{D}m_n$ .

Part (ii). Let  $\hat{a}_k$ ,  $\hat{b}_k$ ,  $a_k$ , and  $b_k$  denote the quantities defined in Appendix B. We write

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

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

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

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

$A$  is  $o_p(1)$  due to two results. First,  $\left( \mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k \right)^{-1}$  is  $O_p(1)$  by Lemma B.3. Second,  $I_1 = K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k$  is  $o_p(1)$  by claim 1 in the proof of part (i).

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

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

where (1) holds by the inequality (B.2) in Appendix B, (2) holds by Lemma B.3 and the Central Limit Theorem, and (3) holds by Lemma B.3.

### A.1.2 Proof of Lemma 2

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

Recall that  $I_1 = I_{1,1} + I_{1,2}$ , where  $I_{1,1}$  and  $I_{1,2}$  are defined in (A.2) and (A.3), respectively. Note that the proof of  $I_{1,2} = o_p(1)$  also applies when  $K_n = O(n)$ ; see derivations in (A.4). Therefore, it is sufficient to show that  $I_{1,1} = o_p(1)$ . Since  $E[I_{1,1}] = 0$ , it is sufficient to show

that  $E[(e_j^\top I_{1,1})^2] = o(1)$ , where  $e_j$  is the  $j$ -th column of  $\mathbb{I}_d$ . Consider the next derivations,

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

where

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

Note that  $I_{1,1,1} = o(1)$  and  $I_{1,1,2} = 0$ . The former holds by Lemma B.1 and Assumption 3.2, while the latter by part (iii) of Assumption 3.1 and the Law of Iterated Expectations.

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

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

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

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

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

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

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

## A.2 Proofs for Section 4

*Proof of Theorem 4.1.* By standard arguments (e.g., Newey and Smith (2004)) and  $\varphi \in (1/4, 1/2)$ , we have  $\sqrt{n} \left( \hat{\theta}_{n, K_n}^{*,(2)} - \theta_0 \right) = \mathcal{T}_n^* + o_p(n^{1/2-2\varphi})$ . Therefore, it is sufficient to show  $\sqrt{n} \left( \hat{\theta}_{n, K_n}^{(2)} - \hat{\theta}_{n, K_n}^{*,(2)} \right) = \mathcal{T}_{n, K_n}^{nl} + o_p(n^{1/2-2\varphi})$ . To show this, we proceed as in the proof of part (ii) in Lemma A.1, and we write

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

Lemma B.4 and  $\varphi \in (1/4, 1/2)$  imply  $B = o_p(n^{1/2-2\varphi})$ . Then, it is sufficient to show  $A = \mathcal{T}_{n, K_n}^{nl} + o_p(n^{1/2-2\varphi})$ . Using the identity (B.1) in Appendix B, we write

$$A = A_1 - (1 + A_2)^{-1} A_1 A_2$$

where  $A_1 = K_n^{-1/2} \sum_{k=1}^{K_n} \hat{a}_k - a_k$  and  $A_2 = n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k$ . Lemmas B.4 and B.5 imply  $A_1 = \mathcal{T}_{n, K_n}^{nl} + o_p(n^{1/2-2\varphi})$  and  $A_1 = O_p(n^{1/2-2\varphi})$  since  $Var[\mathcal{T}_{n, K_n}^{nl}] \propto n^{1/2-2\varphi}$  when  $\mathcal{V} > 0$ , which holds by Assumption 4.2. Lemma B.4 implies  $A_2 = o_p(1)$  and  $(1 + A_2)^{-1} = O_p(1)$ . Then,  $A = \mathcal{T}_{n, K_n}^{nl} + o_p(n^{1/2-2\varphi})$ , which completes the proof of this theorem.  $\square$

*Proof of Theorem 4.2.* It follows by Lemma B.5 and standard calculations.  $\square$

## B Supporting Technical Results

Let  $e_j$  be the  $j$ -column of the identity matrix  $\mathbb{I}_d \in \mathbf{R}^{d \times d}$ . We use the following notation in the proofs of the main results in Appendix A.1:

$$\begin{aligned} \hat{a}_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} G^{-1} \hat{m}_i \\ \hat{b}_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} \left( G^{-1} \hat{\psi}_i^a - \mathbb{I}_d \right) \\ a_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} G^{-1} m_i \\ b_k &= n_k^{-1/2} \sum_{i \in \mathcal{I}_k} \left( G^{-1} \psi_i^a - \mathbb{I}_d \right) \\ D_\eta m_i[\hat{\eta}_i - \eta_i] &= \sum_{t=1}^p (\hat{\eta}_i - \eta_i)^\top \partial_\eta m_t(W_i, \theta_0, \eta_i) \cdot e_t \end{aligned}$$

$$\begin{aligned}
D_\eta \psi_i[\hat{\eta}_i - \eta_i] &= \sum_{s=1}^p \sum_{t=1}^p (\hat{\eta}_i - \eta_i)^\top \partial_\eta \psi_{t,s}(W_i, \theta_0, \eta_i) \cdot e_t e_s^\top \\
D_\eta^2 \tilde{m}_i[\hat{\eta}_i - \eta_i] &= \sum_{t=1}^p (\hat{\eta}_i - \eta_i)^\top \partial_\eta m_t(W_i, \theta_0, \tilde{\eta}_{t,i})(\hat{\eta}_i - \eta_i) \cdot e_t \\
D_\eta^2 \psi_i[\hat{\eta}_i - \eta_i] &= \sum_{s=1}^p \sum_{t=1}^p (\hat{\eta}_i - \eta_i)^\top \partial_\eta \psi_{t,s}(W_i, \theta_0, \tilde{\eta}_{t,s,i})(\hat{\eta}_i - \eta_i) \cdot e_t e_s^\top \\
\mathcal{D}m_n &= \max_{1 \leq k \leq K_n} \left\| n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta m_i[\hat{\eta}_i - \eta_i] \right\| \\
\mathcal{D}\psi_n^a &= \max_{1 \leq k \leq K_n} \left\| n_k^{-1/2} \sum_{i \in \mathcal{I}_k} D_\eta \psi_i^a[\hat{\eta}_i - \eta_i] \right\|,
\end{aligned}$$

where  $\tilde{\eta}_{t,i}$  and  $\tilde{\eta}_{t,s,i}$  are in  $\mathcal{E}$  and exist due Taylor expansions with Lagrange remainder for each  $t$  and  $s$ . We also use the next identity and inequality for square matrices  $\mathbb{M}$  and  $\mathbb{M}_1$ :

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

$$\|(\mathbb{I}_d + \mathbb{M})^{-1} - (\mathbb{I}_d + \mathbb{M}_1)^{-1}\| \leq \|(\mathbb{I}_d + \mathbb{M})^{-1}\| \cdot \|\mathbb{M} - \mathbb{M}_1\| \cdot \|(\mathbb{I}_d + \mathbb{M}_1)^{-1}\| \quad (\text{B.2})$$

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

- (i)  $E[(e_j^\top (D_\eta m_i[\hat{\eta}_i - \eta_i]))^2] \leq CE \left[ \|\hat{\eta}_i - \eta_i\|^2 \right]$  for  $i \in \mathcal{I}_k$  and  $\hat{\eta}_i = \hat{\eta}_k(X_i)$
- (ii)  $E[(e_j^\top (D_\eta m_i[\hat{\eta}_i - \hat{\eta}_i^\ell]))^2] \leq CE \left[ \|\hat{\eta}_i - \hat{\eta}_i^\ell\|^2 \right]$  for  $i \in \mathcal{I}_k$
- (iii)  $\frac{1}{2} \left\| D_\eta^2 \tilde{m}_i[\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \right\| \leq C \|\hat{\eta}_i - \eta_i\|^2$  for  $i \in \mathcal{I}_k$
- (iv)  $\frac{1}{2} \left\| D_\eta^2 \tilde{\psi}_i^a[\hat{\eta}_i - \eta_i, \hat{\eta}_i - \eta_i] \right\| \leq C \|\hat{\eta}_i - \eta_i\|^2$  for  $i \in \mathcal{I}_k$

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

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

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

- (i)  $\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} b_k\right)^{-1} = O_p(1)$
- (ii)  $\left(\mathbb{I}_d + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1} = O_p(1)$
- (iii)  $n^{-1/2} \sum_{i=1}^n \|\hat{\eta}_i - \eta_i\|^2 = o_p(1)$
- (iv)  $n^{-1} \sum_{i=1}^n G^{-1}(\psi^a(W_i, \hat{\eta}_i) - \psi^a(W_i, \eta_i)) = o_p(1)$ .

**Lemma B.4.** *Let Assumptions 3.1, 4.1, and 4.2 hold and let  $K_n$  be such that  $K_n \leq n$ . Then,*

- (i)  $n^{1/2} \sum_{i=1}^n G^{-1}(m(W_i, \theta_0, \hat{\eta}_i) - m(W_i, \theta_0, \eta_i)) = \mathcal{T}_{n, K_n}^{nl} + o_p(n^{1/2-2\varphi})$ .
- (ii)  $n^{-1} \sum_{i=1}^n G^{-1}(\psi^a(W_i, \hat{\eta}_i) - \psi^a(W_i, \eta_i)) = o_p(n^{-\varphi})$ .
- (iii)  $\left(1 + n_k^{-1/2} K_n^{-1} \sum_{k=1}^{K_n} \hat{b}_k\right)^{-1} = O_p(1)$

**Lemma B.5.** *Let Assumptions 3.1, 4.1, and 4.2 hold and let  $K_n$  be such that  $K_n \leq n$ . Then,*

- (i)  $E[\mathcal{T}_{n, K_n}^{nl}] = \mathcal{B} \left(\frac{K_n}{K_n-1}\right)^{2\varphi} n^{1/2-2\varphi} + o(n^{1/2-2\varphi})$ .
- (ii)  $Var[\mathcal{T}_{n, K_n}^{nl}] = \mathcal{V} \left(\frac{K_n^2-3K_n+3}{(K_n-1)^2}\right) \left(\frac{K_n}{K_n-1}\right)^{4\varphi-1} n^{1-4\varphi} + o(n^{1-4\varphi})$ .
- (iii)  $Cov[\mathcal{T}_n^*, \mathcal{T}_{n, K_n}^{nl}] = \frac{1}{2} \mathcal{C} \left(\frac{K_n}{K_n-1}\right)^{2\varphi-1/2} n^{1/2-2\varphi} + o(n^{1/2-2\varphi})$ .

## C Examples

**Example C.1** (Average Treatment Effect). Let  $A \in \{0, 1\}$  denote a binary treatment status,  $Y(a)$  denote the potential outcome under treatment  $a \in \{0, 1\}$ ,  $X$  denote a vector of covariates, and  $Y = AY(1) + (1-A)Y(0)$  denote the observed outcome. The available data is modeled by the vector  $W = (Y, A, X)$ . The parameter of interest is  $\theta_0 = E[Y(1) - Y(0)]$ , which is the expectation of the treatment effect when the treatment is mandated across the entire population, also known as the ATE. Under the selection-on-observables assumption,  $(Y(1), Y(0)) \perp A \mid X$ , the ATE can be identified by a moment condition such as (2.1) using a moment function like (2.2) with  $\psi^a(W, \eta) = 1$  and

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

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

$$\eta_0(X) = (E[Y | X, A = 1], E[Y | X, A = 0], (E[A | X])^{-1}, (E[1 - A | X])^{-1})^\top.$$

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).  $\square$

**Example C.2** (Difference-in-Differences). This example considers the average treatment effect on the treated in difference-in-differences research designs with two periods and panel data. 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  $\theta_0 = E[Y_1(1) - Y_1(0) | A = 1]$ , 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 a conditional parallel trend assumption,  $E[Y_1(0) - Y_0 | X, A = 1] = E[Y_1(0) - Y_0 | X, A = 0]$ , to identify the ATT-DID by a moment condition, such as (2.1), using a moment function like (2.2) with  $\psi^a(W, \eta) = A$  and

$$\psi^b(W, \eta) = A(Y_1 - Y_0 - \eta_1) + (1 - A)(1 - \eta_2)(Y_1 - Y_0 - \eta_1),$$

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

$$\eta_0(X) = (E[Y_1 - Y_0 | X, A = 0], (E[1 - A | X])^{-1})^\top.$$

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

**Example C.3** (Local Average Treatment Effect). 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  $\theta_0 = E[Y(1) - Y(0) | D(1) > D(0)]$ , which is the expected treatment effect for the sub-population that complies with the assigned treatment, also known as

LATE (Imbens and Angrist, 1994). Frölich (2007) and Singh and Sun (2024) used a selection-on-observables assumption,  $(Y(1), Y(0), D(1), D(0)) \perp Z \mid X$ , to identify the LATE by a moment condition, such as (2.1), using a moment function like (2.2) with

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

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

$$\begin{aligned}\eta_0(X) &= (E[Y \mid X, Z = 1], E[Y \mid X, Z = 0], E[D \mid X, Z = 1], \\ &E[D \mid X, Z = 0], (E[Z \mid X])^{-1}, (E[1 - Z \mid X])^{-1})^\top.\end{aligned}$$

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).  $\square$

## References

- AHRENS, A., V. CHERNOZHUKOV, C. HANSEN, D. KOZBUR, M. SCHAFFER, AND T. WIEMANN (2025): “An introduction to double/debiased machine learning,” *arXiv preprint arXiv:2504.08324*.
- AHRENS, A., C. B. HANSEN, M. E. SCHAFFER, AND T. WIEMANN (2024a): “ddml: Double/debiased machine learning in Stata,” *The Stata Journal*, 24, 3–45.
- (2024b): “Model averaging and double machine learning,” *arXiv preprint arXiv:2401.01645*.
- ANDREWS, D. W. (1994): “Asymptotics for semiparametric econometric models via stochastic equicontinuity,” *Econometrica: Journal of the Econometric Society*, 43–72.
- ARGAÑARAZ, F. (2025): “Automatic Debiased Machine Learning of Structural Parameters with General Conditional Moments,” *arXiv preprint arXiv:2512.08423*.
- BACH, P., V. CHERNOZHUKOV, M. S. KURZ, AND M. SPINDLER (2022): “DoubleML—an object-oriented implementation of double machine learning in python,” *Journal of Machine Learning Research*, 23.
- 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.

- BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND K. KATO (2015): “Some new asymptotic theory for least squares series: Pointwise and uniform results,” *Journal of Econometrics*, 186, 345–366.
- BELLONI, A., V. CHERNOZHUKOV, I. FERNANDEZ-VAL, AND C. HANSEN (2017): “Program evaluation and causal inference with high-dimensional data,” *Econometrica*, 85, 233–298.
- BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011): “Square-root lasso: pivotal recovery of sparse signals via conic programming,” *Biometrika*, 98, 791–806.
- BICKEL, P. J. (1982): “On adaptive estimation,” *The Annals of Statistics*, 647–671.
- BICKEL, P. J. AND Y. RITOV (2003): “Nonparametric estimators which can be” plugged-in,” *The Annals of Statistics*, 31, 1033–1053.
- BUGNI, F. A. AND I. A. CANAY (2021): “Testing continuity of a density via g-order statistics in the regression discontinuity design,” *Journal of Econometrics*, 221, 138–159.
- CAI, Y. (2022): “Linear Regression with Centrality Measures,” *arXiv preprint arXiv:2210.10024*.
- CALLAWAY, B. AND P. H. SANT’ANNA (2021): “Difference-in-differences with multiple time periods,” *Journal of econometrics*, 225, 200–230.
- CATTANEO, M. D. AND M. JANSSON (2018): “Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency,” *Econometrica*, 86, 955–995.
- CHANG, N.-C. (2020): “Double/debiased machine learning for difference-in-differences models,” *The Econometrics Journal*, 23, 177–191.
- CHEN, Q., V. SYRGKANIS, AND M. AUSTERN (2022): “Debiased machine learning without sample-splitting for stable estimators,” *Advances in Neural Information Processing Systems*, 35, 3096–3109.
- CHEN, X. (2007): “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, 6, 5549–5632.
- CHENG, X., A. SÁNCHEZ-BECERRA, AND A. J. SHEPHARD (2023): “How to Weight in Moments Matching: A New Approach and Applications to Earnings Dynamics,” *CEMMAP working paper CWP13/23*.

- CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, AND W. NEWEY (2017): “Double/debiased/neyman machine learning of treatment effects,” *American Economic Review*, 107.
- 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., C. HANSEN, N. KALLUS, M. SPINDLER, AND V. SYRGKANIS (2024): “Applied causal inference powered by ML and AI,” *arXiv preprint arXiv:2403.02467*.
- CHERNOZHUKOV, V., W. K. NEWEY, AND R. SINGH (2022b): “Automatic debiased machine learning of causal and structural effects,” *Econometrica*, 90, 967–1027.
- (2022c): “Debiased machine learning of global and local parameters using regularized Riesz representers,” *The Econometrics Journal*, 25, 576–601.
- CHI, C.-M., P. VOSSLER, Y. FAN, AND J. LV (2022): “Asymptotic properties of high-dimensional random forests,” *The Annals of Statistics*, 50, 3415–3438.
- DONALD, S. G. AND W. K. NEWEY (2001): “Choosing the number of instruments,” *Econometrica*, 69, 1161–1191.
- ESCANCIANO, J. AND J. TERSCHUUR (2023): “Machine Learning Inference on Inequality of Opportunity,” .
- ESCANCIANO, J. C. AND T. PÉREZ-IZQUIERDO (2023): “Automatic Locally Robust Estimation with Generated Regressors,” *arXiv preprint arXiv:2301.10643*.
- FARRELL, M. H. (2015): “Robust inference on average treatment effects with possibly more covariates than observations,” *Journal of Econometrics*, 189, 1–23.
- FARRELL, M. H., T. LIANG, AND S. MISRA (2021): “Deep neural networks for estimation and inference,” *Econometrica*, 89, 181–213.
- (2025): “Deep learning for individual heterogeneity: An automatic inference framework,” *arXiv preprint arXiv:2010.14694*.

- FRÖLICH, M. (2007): “Nonparametric IV estimation of local average treatment effects with covariates,” *Journal of Econometrics*, 139, 35–75.
- HAHN, J. (1998): “On the role of the propensity score in efficient semiparametric estimation of average treatment effects,” *Econometrica*, 315–331.
- HAHN, J. AND G. RIDDER (2013): “Asymptotic variance of semiparametric estimators with generated regressors,” *Econometrica*, 81, 315–340.
- 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. NEWWEY (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*.
- LINTON, O. (1995): “Second order approximation in the partially linear regression model,” *Econometrica: Journal of the Econometric Society*, 1079–1112.
- LIU, Y. AND F. MOLINARI (2025): “Inference for an Algorithmic Fairness-Accuracy Frontier,” *arXiv preprint arXiv:2402.08879*.
- LIU, Y., F. MOLINARI, AND A. VELEZ (2026): “Identification and Inference for Algorithmic Frontiers with Selective Labels,” *Work in progress*.
- NEWWEY, W. K. (1990): “Efficient instrumental variables estimation of nonlinear models,” *Econometrica: Journal of the Econometric Society*, 809–837.
- (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica: Journal of the Econometric Society*, 1349–1382.
- (1997): “Convergence rates and asymptotic normality for series estimators,” *Journal of econometrics*, 79, 147–168.

- NEWWEY, W. K. AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” *Handbook of econometrics*, 4, 2111–2245.
- NEWWEY, W. K. AND J. R. ROBINS (2018): “Cross-fitting and fast remainder rates for semiparametric estimation,” *arXiv preprint arXiv:1801.09138*.
- NEWWEY, W. K. AND R. J. SMITH (2004): “Higher order properties of GMM and generalized empirical likelihood estimators,” *Econometrica*, 72, 219–255.
- NOACK, C., T. OLMA, AND C. ROTHE (2024): “Flexible covariate adjustments in regression discontinuity designs,” *arXiv preprint arXiv:2107.07942*.
- PARK, G. (2024): “Debiased Machine Learning when Nuisance Parameters Appear in Indicator Functions,” *arXiv preprint arXiv:2403.15934*.
- RAFI, A. (2023): “Efficient semiparametric estimation of average treatment effects under covariate adaptive randomization,” *arXiv preprint arXiv:2305.08340*.
- ROBINS, J. M. AND A. ROTNITZKY (1995): “Semiparametric efficiency in multivariate regression models with missing data,” *Journal of the American Statistical Association*, 90, 122–129.
- ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American statistical Association*, 89, 846–866.
- ROBINSON, P. M. (1988): “Root-N-consistent semiparametric regression,” *Econometrica: Journal of the Econometric Society*, 931–954.
- ROTHE, C. AND S. FIRPO (2019): “Properties of doubly robust estimators when nuisance functions are estimated nonparametrically,” *Econometric Theory*, 35, 1048–1087.
- ROTHENBERG, T. J. (1984): “Approximating the distributions of econometric estimators and test statistics,” *Handbook of econometrics*, 2, 881–935.
- SANT’ANNA, P. H. AND J. ZHAO (2020): “Doubly robust difference-in-differences estimators,” *Journal of econometrics*, 219, 101–122.
- SCHARFSTEIN, D. O., A. ROTNITZKY, AND J. M. ROBINS (1999): “Adjusting for non-ignorable drop-out using semiparametric nonresponse models,” *Journal of the American Statistical Association*, 94, 1096–1120.

- SCHMIDT-HIEBER, J. (2020): “Nonparametric regression using deep neural networks with ReLU activation function,” *The Annals of Statistics*, 48, 1875 – 1897.
- SEMENOVA, V. (2023a): “Adaptive estimation of intersection bounds: a classification approach,” *arXiv preprint arXiv:2303.00982*.
- (2023b): “Debiased machine learning of set-identified linear models,” *Journal of Econometrics*, 235, 1725–1746.
- SEMENOVA, V. AND V. CHERNOZHUKOV (2021): “Debiased machine learning of conditional average treatment effects and other causal functions,” *The Econometrics Journal*, 24, 264–289.
- SINGH, R. AND L. SUN (2024): “Double robustness for complier parameters and a semi-parametric test for complier characteristics,” *The Econometrics Journal*, 27, 1–20.
- TAN, Z. (2006): “Regression and weighting methods for causal inference using instrumental variables,” *Journal of the American Statistical Association*, 101, 1607–1618.
- TIBSHIRANI, R. (1996): “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.
- VAN DE GEER, S. A. (2008): “High-dimensional generalized linear models and the lasso,” .