Discussion Paper Series

IZA DP No. 18438

March 2026

# An Introduction to Double/ Debiased Machine Learning

**Achim Ahrens**
CERGE-EI, University of Lausanne and Immigration Policy Lab

**Victor Chernozhukov**
Massachusetts Institute of Technology

**Christian Hansen**
University of Chicago Booth School of Business

**Damian Kozbur**
University of Zurich

**Mark E. Schaffer**
Heriot-Watt University and IZA@LISER

**Thomas Wiemann**
University of Chicago Booth School of Business




ISSN: 2365-9793

IZA DP No. 18438 March 2026

# An Introduction to Double/Debiased Machine Learning*

## Abstract

This paper provides an introduction to Double/Debiased Machine Learning (DML). DML is a general approach to performing inference about a target parameter in the presence of nuisance functions: objects that are needed to identify the target parameter but are not of primary interest. Nuisance functions arise naturally in many settings, such as when controlling for confounding variables or leveraging instruments. The paper describes two biases that arise from nuisance function estimation and explains how DML alleviates these biases. Consequently, DML allows the use of flexible methods, including machine learning tools, for estimating nuisance functions, reducing the dependence on auxiliary functional form assumptions and enabling the use of complex non-tabular data, such as text or images. We illustrate the application of DML through simulations and empirical examples. We conclude with a discussion of recommended practices. A companion website includes additional examples and references to other resources.

## JEL classification

C14, C21, C23, C26

## Keywords

causal inference, econometrics, high-dimensional models, machine learning, nonparametric estimation

## Corresponding author

Mark E. Schaffer
m.e.schaffer@hw.ac.uk

* We thank Thiago Cacicedo dos Santos, Francesca Codega, Sara Drango, Teresa Freitas Monteiro, Samuel Higbee, Štěpán Jurajda, Rafael Lalive, Hugo Lopez, Claudia Marangon, Moritz Marbach, Max Maydanchik, Eoin McLaughlin, and Alessandra Stampi-Bombelli for reading and providing helpful comments on drafts of the paper. We also thank seminar participants at the Luxembourg Institute of Socio-Economic Research (LISER) and participants at the *Tools, Data, and Methods Workshop* at the University of Exeter. We further thank the editor, David Romer, and five anonymous referees for constructive comments and suggestions. Computational resources were provided by ETH Zürich's DeSciL, and the e-INFRA CZ project. Achim Ahrens acknowledges support from the project New Technologies and Changes in Education, Research and Labor Market (reg. no. CZ.02.01.01/00/23_025/0008693), cofunded by the European Union. We used OpenAI's GPT-4o, GPT-5, GPT-5.2 and Anthropic's Claude Opus 4.5 to assist with proofreading and editing during the preparation of this manuscript. All errors are our own. The regularly updated companion website dmlguide.github.io includes replication files, an expanding set of additional examples and discussion, and links to other resources.

# 1 Introduction

A large share of empirical research in economics aims to provide insights into the statistical relationships among two or more variables. For example, a common research goal is to understand the causal impact of a policy on economic outcomes. Target parameters summarizing these relationships, including average treatment effects or regression coefficients, frequently depend on *nuisances*—auxiliary objects that must be accounted for to identify the parameter of interest but are not themselves of primary interest, such as regression coefficients on control variables.

As a concrete example that we will revisit in our empirical illustrations, consider Dube et al. (2020), who study monopsony power on the online platform MTurk using partially linear regression:

$$Y = \theta_0 D + g_0(X) + \varepsilon$$

where $Y$ is the logarithm of the time it takes for a posted job to be filled, $D$ is the logarithm of the reward of the job, $X$ denotes observed features of tasks including their type and complexity, and $\varepsilon$ is assumed to be uncorrelated with $D$ and mean independent of $X$, i.e., $\mathrm{E}[D\varepsilon] = 0$ and $\mathrm{E}[\varepsilon|X] = 0$. Their target parameter is the regression coefficient $\theta_0$, which they interpret as a measure of the negative labor supply elasticity.

In this example, $g_0(\cdot)$ is a nuisance function. We are not primarily interested in how task features, outside of reward, relate to the outcome. However, Dube et al. (2020) emphasize that task heterogeneity needs to be accounted for to meaningfully interpret $\theta_0$.

If $g_0(\cdot)$ were known, estimation of $\theta_0$ could proceed by regressing $Y - g_0(X)$ onto $D$, which is equivalent to estimation based on the moment condition

$$\mathrm{E}[m(W; \theta_0, g_0)] = \mathrm{E}[(Y - g_0(X) - D\theta_0)D] = 0$$

where $W = (Y, D, X)$ denote observed random variables. Of course, $g_0(\cdot)$ will typically not be known. A common strategy to simplify the problem is to *assume* that $g_0(X) = X'\beta_0$ with unknown coefficient $\beta_0$, in which case the model reduces to the familiar multiple regression model. When the dimension of $X$ is much smaller than the sample size, estimation would then proceed by ordinary least squares (OLS) of $Y$ on $D$ and $X$.

Even in the simple multiple regression model, we have a nuisance parameter, $\beta_0$. Conventional regression estimates it jointly with the target parameter $\theta_0$. Alternatively, one can partial $X$ out from both $D$ and $Y$ to obtain residuals that isolate the variation identifying $\theta_0$. By the Frisch-Waugh-Lovell Theorem, regressing these residuals on each other yields an estimate of $\theta_0$ that is numerically equivalent to that obtained from regressing $Y$

on $D$ and $X$. This notion of "partialling out" is related to a broader principle for handling nuisance parameters that we emphasize throughout the review.

In the actual Dube et al. (2020) example, some of the task characteristics are captured as text data, which makes it difficult to justify *ad hoc* parametric assumptions like those in the linear model. Instead, Dube et al. (2020) allow $g_0(\cdot)$ to be a flexible, *high-dimensional* function rather than committing to a low-dimensional functional form. In contrast to the low-dimensional linear case, where $g_0(X) = X'\beta_0$ reduces to estimating a small set of coefficients, estimation of $g_0(\cdot)$ must then also accommodate rich nonlinear variation.

This regression setup illustrates a broader template common in empirical research, in which the target parameter $\theta_0$ is defined as the solution to a moment condition:

$$\theta_0 : \ \mathrm{E}[m(W; \theta_0, \eta_0)] = 0. \tag{1}$$

Here, $m$ is a score (or moment) function and $W$ again denotes observed random variables. The parameter $\eta_0$ denotes a nuisance object, which is not of direct interest but is used to define $\theta_0$. $\eta_0$ is often high-dimensional. For example, $\eta_0$ will represent a vector of conditional expectation functions in many interesting cases. This *semi-parametric* structure encompasses the regression example above, where the target parameter is the coefficient on $D$. It more generally applies to the estimation of many other canonical parameters, including average treatment effects, parameters in linear instrumental variables models, local average treatment effects, dynamic treatment effects in staggered adoption designs, and parameters in nonlinear structural models.

High-dimensional nuisance parameters can arise in several ways: (i) when the nuisance function depends on only a few covariates, controls, or instruments, but no parametric model is specified; (ii) when there are many such variables, even under parametric assumptions; or (iii) when numerous variables enter through unknown functions. High-dimensionality is increasingly common in applications using text or image data (Gentzkow, Kelly, and Taddy, 2019), but it can also arise in simpler settings. Even a single continuous covariate may create a high-dimensional problem—for example, when identification relies on rainfall instruments that can be nonlinearly related to an endogenous regressor (e.g., Hidalgo et al., 2010; Gilchrist and Sands, 2016; Dustmann, Fasani, and Speciale, 2017).

As rich data become more common in applied research, there is growing appreciation that traditional functional form assumptions are often difficult to justify, motivating the use of more flexible tools for estimating nuisance parameters. As a result, there is increasing interest in using machine learning (ML) methods, which provide flexible tools for estimating high-dimensional nuisance parameters. A natural use of ML would then be to obtain nuisance parameter estimates, $\widehat{\eta}$, and use these in place of $\eta_0$ in (1). Specifically,

we can define an estimator $\widehat{\theta}$ of $\theta_0$ as a solution to the sample analog of (1):

$$\widehat{\theta}: \frac{1}{n}\sum_{i=1}^{n} m(W_i; \widehat{\theta}, \widehat{\eta}) = 0$$

where $W_i$ denote observed variables for observations $i = 1, ..., n$. However, the resulting "plug-in" estimator $\widehat{\theta}$—so-called because it "plugs" $\widehat{\eta}$ in for $\eta_0$—can behave poorly and lead to misleading conclusions due to errors in estimating $\eta_0$ propagating into $\widehat{\theta}$.

In settings with high-dimensional nuisance parameters, standard asymptotic approximations may fail due to two distinct forms of sensitivity to nuisance parameter estimation, termed *regularization bias* and *overfitting bias*. Both terms describe channels through which using an estimated nuisance parameter, $\widehat{\eta}$, instead of the true but unknown $\eta_0$, can distort the behavior of the plug-in estimator $\widehat{\theta}$. A leading manifestation of these distortions is bias, as the terminology suggests, but they more broadly invalidate conventional inference methods that fail to account for nuisance estimation.

Roughly speaking, regularization bias refers to the direct impact of estimation error in $\widehat{\eta}$ on the plug-in estimator $\widehat{\theta}$ that results from the difference between $m(W; \theta_0, \widehat{\eta})$ and $m(W; \theta_0, \eta_0)$. Overfitting bias refers to a more subtle issue. Because $\widehat{\eta}$ is an estimator, it is itself a random function of the data. $\widehat{\eta}$ is thus generally correlated with the observations $\{W_i\}_{i=1}^n$ also used in the estimating equation $\frac{1}{n}\sum_{i=1}^n m(W_i; \theta, \widehat{\eta})$. When this dependence is strong, for example due to "overfitting", it may generate large differences between $\frac{1}{n}\sum_{i=1}^n m(W_i; \theta, \widehat{\eta})$ and $\frac{1}{n}\sum_{i=1}^n m(W_i; \theta, \eta_0)$, which results in poor performance of $\widehat{\theta}$.

Both regularization and overfitting biases are major concerns in high-dimensional contexts. Accurately estimating high-dimensional $\eta_0$ is inherently difficult and often results in non-negligible errors. Further, estimation in high-dimensional settings typically relies on highly data-adaptive procedures—such as modern ML methods—which amplify the risk of overfitting bias. As such, mitigating these biases is the focus of a large and rapidly growing literature in statistics and econometrics that builds from classic ideas in semiparametric and nonparametric estimation. One method that provides a solution in a wide variety of empirical settings, and that is the topic of this review, is double/debiased machine learning (henceforth DML; Chernozhukov et al., 2018).

DML provides a blueprint for alleviating both regularization and overfitting bias. At its core, DML combines two classical ideas from the rich literature on semiparametric inference—using *Neyman orthogonal scores*[1] to alleviate regularization bias and using *cross-fitting* to alleviate overfitting bias—in a common methodological framework. Ney-

[1]Such scores are also referred to as orthogonal scores, orthogonal moments, locally robust moments, debiased moments, influence functions, and pathwise derivatives. We follow Chernozhukov et al. (2018) and use the term "Neyman orthogonal scores" in homage to Neyman's early contributions, e.g., Neyman (1959) and Neyman (1979).

man orthogonality ensures that plugging in estimates that are close to, but not exactly equal to, $\eta_0$ does not lead to large changes in the moment condition (1).[2] Cross-fitting, a form of sample splitting, alleviates potential dependence between nuisance estimates $\widehat{\eta}$ and parts of the data used for estimating the target parameter. Used in conjunction, DML's two core components significantly reduce the impact of nuisance estimation on estimates of the target parameter. However, high-quality estimation of the target parameter $\theta_0$ still requires nuisance parameters to be estimated sufficiently well. Consequently, theoretical results for DML assume specific convergence rate conditions on nuisance estimators. Many estimators, including ML methods, can satisfy these conditions.

From a practical perspective, DML enables researchers to leverage a wide range of ML tools, making it particularly valuable in complex data scenarios involving numerous variables, images, or text data. Importantly, the benefits of ML also extend to traditional research settings with fewer covariates and conventional tabular data as they remove the need for researchers to commit beforehand to specific parametric (often linear) models. Furthermore, DML is easy to implement, applicable to a wide range of econometric settings, and readily available in existing software packages, including Stata, R, and Python (e.g., Bach et al., 2021; Bach et al., 2022; Ahrens et al., 2024; Wiemann et al., 2023). DML thus has the potential to enhance the credibility of research findings in a broad spectrum of settings, either when used as a complementary robustness check or when the application necessitates the use of flexible estimation methods for nuisance objects.

While DML offers a framework for combining flexible nuisance estimation with valid asymptotic inference, its implementation raises important challenges. Available theoretical results assume the nuisance functions are estimated with sufficiently high accuracy. Achieving these convergence rates for modern ML methods often demands strong assumptions and special tuning, and they may not hold for off-the-shelf algorithms. These theoretical qualifications manifest in practical problems where empirical results depend on implementation choices, such as selecting and tuning an ML method for nuisance estimation. Assessing the quality of nuisance estimators is often difficult and, in some applications, different seemingly reasonable choices can lead to substantively different conclusions. This paper therefore aims not only to motivate and explain DML, but also to guide its application in empirical research, emphasizing the need for careful diagnostic analysis and robustness checks. To this end, we divide our review into two parts.

First, in Sections 2 and 3, we introduce the DML blueprint at a high level. Section 2 discusses the practical implications of nuisance estimation and the role of DML's two key components in their remedy. Section 3 summarizes the asymptotic properties of DML

[2]Neyman orthogonality is not guaranteed for all scores that serve to identify a parameter of interest. We show how to construct a Neyman orthogonal score from a given score in Appendix B.

and contains algorithmic details on implementation of generic DML estimators.

In the second part of the paper, we turn to simulations and empirical applications, found in Section 4 through Section 6. These examples illustrate DML and provide discussion of key issues that arise in its implementation.

In Section 4, we present results from two simulation examples. The first is a simple linear IV example that demonstrates the importance of cross-fitting. The second compares the benefits of the DML average treatment effect estimator with inverse propensity weighted and regression-based estimators.

In Section 5, we illustrate how DML can be leveraged to reduce the dependence on functional forms in staggered adoption designs with covariates. We revisit the analysis of Dobkin et al. (2018), who study the economic consequences of hospital admission. We estimate group-time average treatment effects on the treated under a conditional parallel trends assumption, and show how DML inference applies to dynamic average treatment effects. We note that cross-fitting introduces an additional source of randomness induced by sample-splitting. Part of our aim in this example is to illustrate a simple approach to aid in gauging the impact of this source of randomness.

In Section 6, we apply DML to estimation of regression coefficients in the presence of complex covariates. We specifically revisit Dube et al. (2020) who apply DML to estimate the labor supply elasticity in online labor markets using textual controls. We have two main goals in this section: to illustrate the use of complex non-tabular data and, more importantly, to illustrate that DML estimates can vary substantially across otherwise reasonable choices of machine *learners* (i.e., algorithms used to estimate nuisance functions). This sensitivity can lead to qualitatively different conclusions about economically meaningful parameters. Because it seems difficult to know *ex ante* exactly which learner one should choose in these situations, we use this example to discuss robustness checks and suggest strategies for selecting ML algorithms.

Section 7 concludes by summarizing takeaways, raising some caveats, and pointing to potential directions for further research.

To complement the present article, we provide additional resources on our regularly updated website dmlguide.github.io. The materials include replication files, additional examples with code, references to DML software packages, and links to other resources.

In terms of scope, we emphasize that DML—or any other estimation framework—cannot replace careful reasoning about economic parameters and identifying assumptions. Rather, with a well-defined target parameter and corresponding identifying assumptions, DML can aid in obtaining estimates of the target parameter in the presence of complex data structures and without relying on pre-specified functional form assumptions. In other words, DML is useful only after a target parameter is defined and the assumptions

linking observed data to that parameter are well understood.[3] With this understanding, we discuss identification assumptions only with the aim of illustrating the economic content of the applications. Further, we will not review specific ML methods. Varian (2014), Mullainathan and Spiess (2017), Athey and Imbens (2019), and Dell (2024) provide reviews of ML methods targeted at economists. Hastie, Tibshirani, and Friedman (2009) and James et al. (2023) are classic textbook treatments of popular ML methods.

***Literature.*** Inference about low-dimensional target parameters in the presence of high-dimensional or nonparametric nuisance components has a long history in econometrics and statistics. Classic reviews such as Newey and McFadden (1994), Yatchew (1998), Li and Racine (2006), Chen (2007), and Ichimura and Todd (2007) synthesize early work on semiparametric and nonparametric methods, emphasizing how nuisance parameters can be accommodated without fully specifying the data-generating process. More recent surveys shift the focus toward the use of modern machine learning tools for nuisance estimation and their implications for inference; see, e.g., Chernozhukov et al. (2018), Díaz (2020), Hines et al. (2022), and Kennedy (2023a). Our review complements these contributions by emphasizing the practical consequences of nuisance estimation choices for applied empirical work.

DML combines two ideas with deep roots in the semiparametric inference literature: Neyman orthogonal scores and sample splitting. Neyman (1959) introduced orthogonal scores in the context of efficient parametric hypothesis testing. Orthogonal scores later played a central role in the development of modern semiparametric estimation, especially in settings with high-dimensional or nonparametric nuisance parameters. Key contributions developing these ideas include van der Vaart (1991), Andrews (1994), and Newey (1994), with a comprehensive textbook treatment provided by van der Vaart (1998).

Sample splitting has also played a long-standing role in semiparametric inference. It appears in several early contributions to semiparametric estimation; see, for example, Hasminskii and Ibragimov (1978), Bickel (1982), Pfanzagl (1982), Schick (1986), and Bickel and Ritov (1988). In economics, sample splitting has long been used in instrumental variable estimation to mitigate bias from many instruments. See, for instance, Angrist and Krueger (1995b) and Angrist, Imbens, and Krueger (1999) for foundational work and Chao et al. (2012), Hansen and Kozbur (2014), and Chyn, Frandsen, and Leslie (2024) for current developments.

More recently, sample splitting and variations such as cross-fitting have gained renewed attention in high-dimensional contexts. A growing literature shows how these techniques can mitigate problems introduced by overfitting and improve inference when modern ML

[3]This treatment parallels Heckman and Vytlacil (2007), who stress that estimation plays a limited role relative to defining a target parameter and articulating the assumptions that connect it to the data.

methods are used for nuisance estimation. See, for example, Robins et al. (2008), Belloni, Chernozhukov, and Hansen (2010), Belloni et al. (2012), Fan, Guo, and Hao (2012), Robins et al. (2013), Hubbard, Kherad-Pajouh, and van der Laan (2016), Robins et al. (2017), Wager and Athey (2018), Athey et al. (2019), and Athey and Wager (2021).

DML is also related to targeted maximum likelihood (or minimum loss) estimation, which was introduced in Scharfstein, Rotnitzky, and Robins (1999) for treatment effect estimation and generalized by van der Laan and Rubin (2006); see also van der Laan and Rose (2011). Zheng and van der Laan (2011) discuss benefits of sample splitting for targeted maximum likelihood learning. Díaz (2020) expands on the difference between DML and targeted maximum likelihood estimation.

Our focus in this review is on the practical implications of nuisance estimation and the core ideas that motivate DML. Accordingly, we do not attempt a comprehensive survey of the rapidly expanding literature that extends DML in a wide range of directions. We nevertheless highlight several representative strands of this work.

A number of papers adapt DML to canonical empirical settings, including panel data and difference-in-differences designs (Chang, 2020; Chiang et al., 2022; Klosin and Vilgalys, 2023; Abadie et al., 2024; Haddad, Huber, and Zhang, 2024; Clarke and Polselli, 2024; Chiang et al., 2026). Related contributions study the use of DML in instrumental variables and proxy control settings (Jung, Tian, and Bareinboim, 2021; Deaner, 2023; Singh and Sun, 2024). A growing body of research also examines treatment effect and policy parameters, including incremental and dynamic treatment effects, nonparametric policy learning, and localized estimands that depend on complex nuisance components (Bonvini et al., 2021; Lewis and Syrgkanis, 2021; Klosin, 2021; Nie and Wager, 2021; Semenova and Chernozhukov, 2021; Colangelo and Lee, 2023; Foster and Syrgkanis, 2023; Kennedy, 2023b; Sasaki and Ura, 2023; Kallus, Mao, and Uehara, 2024). Other extensions address specific econometric complications. These include settings with generated regressors (Escanciano and Pérez-Izquierdo, 2023), partial or set identification (Semenova, 2023), and sample selection (Bia, Huber, and Lafférs, 2024). Complementing these application-driven contributions, several papers develop general frameworks for the automatic construction of Neyman orthogonal moments for broad classes of target parameters (Chernozhukov et al., 2021; Farrell, Liang, and Misra, 2021a; Chernozhukov et al., 2022; Chernozhukov, Newey, and Singh, 2022a; Chernozhukov, Newey, and Singh, 2022b).

More broadly, DML allows researchers to avoid auxiliary parametric assumptions. While these assumptions simplify estimation, they are seldom motivated by economics and can be detrimental for applications that aim to estimate causal parameters. This perspective connects DML to the recent literature highlighting that statistically convenient estimands, often based on linear models, may fail to even approximate causal effects.

Such failures have been documented in difference-in-differences settings (de Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Baker, Larcker, and Wang, 2022; de Chaisemartin and d'Haultfoeuille, 2023; Roth et al., 2023; Borusyak, Jaravel, and Spiess, 2024). Related concerns arise in linear regression with multiple treatments (Goldsmith-Pinkham, Hull, and Kolesár, 2024) and in instrumental variables settings (Blandhol et al., 2022). By allowing for flexible estimation of nuisance parameters, DML provides a framework for inference that obviates the need for convenient but potentially detrimental parametric assumptions.

# 2 Key Ingredients of DML

This section describes the two essential components that define DML: Neyman orthogonality and cross-fitting. Together, they help control the sensitivity of the target estimator to nuisance estimation, which can substantially improve both the reliability of point estimates and the quality of conventional asymptotic approximations. By alleviating this sensitivity, DML further opens the door for researchers to use a wide range of flexible estimators, including many modern machine learners, for estimating nuisance parameters.

## 2.1 A Semiparametric Framework for DML

We frame our discussion of DML within a relatively general semiparametric framework. There are two key elements of the framework. First, we have a target parameter of interest, $\theta_0$, which is low-dimensional; e.g., $\theta_0$ may be an average treatment effect or a fixed vector of regression coefficients. Second, we have a nuisance parameter $\eta_0$ which may be high-dimensional and potentially complex. In many examples, $\eta_0$ is a vector of conditional expectation functions, such as outcome regressions and propensity scores, though it may take other forms.

Throughout this review, we focus on the case where we observe an *i.i.d.* sample $\{W_i : i = 1, \ldots, n\}$ from a random vector $W$. Each $W_i$ collects the variables relevant for individual $i$. For example, $W_i$ might include an outcome $Y_i$, a treatment variable $D_i$, a vector of controls $X_i$, and excluded instruments $Z_i$. This structure also extends to cross-sectional and panel settings with arbitrary temporal dependence and fixed $T$.[4]

We assume the target parameter is identified by moment conditions

$$\mathrm{E}\left[m(W; \theta_0, \eta_0)\right] = 0, \tag{2}$$

[4]To account for cluster dependence, one may simply redefine $W_i$ to include the data of the $i$th individual over multiple time periods—e.g., $W_i = (Y_{i,t}, D_{i,t}, X_{i,t})_{t=1}^{T}$.

where $m(\cdot;\theta,\eta)$ is a known score function indexed by $\theta$ and nuisance parameter $\eta$ with true values $\theta_0$ and $\eta_0$. We focus exclusively on the case where the score function $m(\cdot;\theta,\eta)$ defines as many constraints as we have parameters of interest, but note that the framework extends to other settings such as GMM as discussed, e.g., in Chernozhukov et al. (2018). Throughout, we assume that the target parameter is strongly identified in the sense that (2) has a unique solution and satisfies regularity conditions such that $\sqrt{n}$-consistent and asymptotically normal inference for $\theta_0$ would be achievable if $\eta_0$ were known.[5]

This semiparametric framework captures a large range of common parameters of interest in empirical research. We discuss four illustrative examples below.

**Example 1. Linear Regression Coefficient.** Consider linear regression with a single variable of interest $D$ and a $p \times 1$ vector of controls $X$ that may include a constant:

$$Y = \theta_0 D + X'\beta_0 + \varepsilon, \quad \mathrm{E}[D\varepsilon] = 0, \quad \mathrm{E}[X\varepsilon] = 0_p \tag{3}$$

where $0_p$ denotes a $p \times 1$ vector of zeros. The coefficient $\theta_0$ on $D$ is the target parameter. The vector of coefficients $\beta_0$ on controls $X$ is the nuisance parameter.

The traditional textbook approach is to estimate both $\theta_0$ and $\beta_0$ by applying OLS to equation (3). This problem can be framed as a semiparametric estimation task by explicitly targeting $\theta_0$ separately from the nuisance parameters. In the linear regression example, this corresponds to another textbook approach: partialling out.

For any random variable $A$, let $\eta_{A,0} = \arg\min_\eta \mathrm{E}[(A - X'\eta)^2]$ denote the best linear predictor coefficient of $A$ given $X$. This definition implies the orthogonality condition

$$\mathrm{E}[X(A - X'\eta_{A,0})] = 0_p. \tag{4}$$

A valid score for $\theta_0$ is then

$$m_{LM}(W;\theta,\eta) = [(Y - X'\eta_Y) - \theta(D - X'\eta_D)]\,(D - X'\eta_D), \tag{5}$$

where the nuisance parameter is $\eta = (\eta_Y', \eta_D')'$ with true value $\eta_0 = (\eta_{Y,0}', \eta_{D,0}')'$. By (3) and the orthogonality condition, (4), $\mathrm{E}[m_{LM}(W;\theta_0,\eta_0)] = 0$.

Equation (5) is the population moment condition underlying the partialling out interpretation of linear least squares regression. It corresponds to projecting $Y$ and $D$ onto $X$ and then estimating $\theta_0$ from a regression using the resulting residuals.

By Frisch-Waugh-Lovell, this score yields the same estimator as OLS of $Y$ on $(D, X)$ in the low-dimensional linear setting. The value of writing the problem in this way is

[5]Extension to weakly identified settings is possible as in, e.g., Chernozhukov, Hansen, and Spindler (2015) and Ma (2023).

therefore primarily conceptual. It makes explicit the construction of a score for the target parameter $\theta_0$ by projecting onto covariates $X$ and appropriate partialling out. This familiar approach generalizes and lies at the core of the key "Neyman orthogonality" property—to be discussed in Section 2.3—that is fundamental for DML. We verify Neyman orthogonality of the score (5) in Appendix A, and discuss a more general construction of Neyman orthogonal scores via "partialling out" in Appendix B.

□

**Example 2. Partially Linear Regression Coefficient.** As discussed in the Introduction, partially linear regression (PLR),

$$Y = \theta_0 D + g_0(X) + \varepsilon, \quad \mathrm{E}[D\varepsilon] = \mathrm{E}[\varepsilon|X] = 0, \tag{6}$$

is a natural, flexible generalization of multiple linear regression.[6]

There are several moment conditions for identifying the PLR coefficient $\theta_0$. Two leading examples are based on the score functions

$$m_{naive}(W;\theta,\eta) = (Y - g(X) - \theta D)D, \tag{7}$$

$$m_{PLM}(W;\theta,\eta) = [(Y - \ell(X)) - \theta(D - r(X))]\,(D - r(X)), \tag{8}$$

where the nuisance parameters are $\eta(X) = g(X)$ in (7) with true value $g_0(X)$, and $\eta(X) = (\ell(X), r(X))$ in (8) with true values $\ell_0(X) = \mathrm{E}[Y|X]$ and $r_0(X) = \mathrm{E}[D|X]$.

The first score is equivalent to regressing $Y - g(\cdot)$ against $D$. The second score corresponds to a "partialling out" approach where both $Y$ and $D$ are residualized with respect to $X$ before regressing the residuals on each other. The latter mirrors the Frisch–Waugh–Lovell logic from linear regression but now allows $X$ to enter flexibly. Note that (8) corresponds to the treatment of the partially linear model in Robinson (1988). While both scores identify the target parameter $\theta_0$, only $m_{PLM}$ satisfies the key "Neyman orthogonality" property that is fundamental for DML. We verify this in Section 2.3. □

**Example 3. Linear IV Coefficient.** In linear instrumental variable (IV) models, it is often unclear how best to use instruments. For instance, when rainfall or weather variables are employed as instruments, researchers face choices such as whether to use rainfall in levels, logarithms, squared terms, or deviations from historical averages (e.g., Hidalgo et al., 2010; Gilchrist and Sands, 2016; Dustmann, Fasani, and Speciale, 2017).

[6] Analogous to linear regression, PLR can be motivated in similar manner as the best *partially* linear approximation to the conditional expectation function. Arguments for economic interest in the best "fully" linear approximation to the conditional expectation function as outlined, e.g., in Angrist and Pischke (2009, Ch. 3), also make PLR an attractive baseline choice in many economic analyses.

Formally, consider the linear structural equation $Y = \theta_0 D + \varepsilon$ where $D$ is an endogenous variable, $Z$ is a vector of excluded instruments, $\mathrm{E}[\varepsilon|Z] = 0$ holds, and we abstract from other covariates. A natural score function is then

$$m_{IV}(W; \theta, \eta) = (Y - \theta D)\eta(Z) \tag{9}$$

where the true value of the nuisance function is $\eta_0(Z) = \mathrm{E}[D|Z]$, which corresponds to the optimal instrument under homoskedasticity. $\mathrm{E}[m_{IV}(W; \theta_0, \eta_0)] = 0$ then follows immediately from the exclusion restriction $\mathrm{E}[\varepsilon|Z] = 0$.[7]

The IV score satisfies the key "Neyman orthogonality" property that is fundamental for DML. We verify this in Appendix A. □

**Example 4. Average Treatment Effect.** A central policy-relevant parameter is the average treatment effect (ATE) of a binary treatment $D$ on an outcome $Y$ defined as

$$\theta_0 = \mathrm{E}\left[Y(1) - Y(0)\right] \tag{10}$$

where $Y(d)$ is the potential outcome under treatment status $d \in \{0, 1\}$.

In non-experimental settings, identification of the ATE relies on two standard conditions: overlap and unconfoundedness (e.g., Imbens and Rubin, 2015). Overlap requires that the probability of treatment is bounded away from 0 and 1 across all covariate values: $0 < \Pr(D = 1|X = x) < 1$ for all $x$. That is, we should see treatment and control observations at all values of $X$. Unconfoundedness requires that the treatment status is independent of potential outcomes after conditioning on the covariates: $(Y(1), Y(0)) \perp D|X$. That is, treatment is as good as randomly assigned after conditioning on $X$.

Under these assumptions, the ATE can be identified using moment conditions. We consider two commonly used scores, the inverse propensity weighted (IPW) score and the augmented IPW (AIPW) score (Newey, 1994; Robins, Rotnitzky, and Zhao, 1994):

$$m_{IPW}(W; \theta, \alpha) = \alpha(D, X)Y - \theta, \tag{11}$$

$$m_{AIPW}(W; \theta, \eta) = \alpha(D, X)(Y - \ell(D, X)) + \ell(1, X) - \ell(0, X) - \theta, \tag{12}$$

where $W = (Y, D, X)$. The true value of the nuisance parameters are $\alpha_0(D, X) = \frac{D}{r_0(X)} - \frac{(1-D)}{1-r_0(X)}$, $r_0(X) = \mathrm{E}[D|X]$, and $\ell_0(D, X) = \mathrm{E}[Y|D, X]$. Under overlap and unconfoundedness, it can be shown that both $\mathrm{E}[m_{IPW}(W; \theta_0, \eta_0)] = 0$ and $\mathrm{E}[m_{AIPW}(W; \theta_0, \eta_0)] = 0$.

[7] Under mean independence, any function $g(Z)$ serves as a valid instrument in the sense of satisfying the moment condition $\mathrm{E}[(Y - \theta_0 D)g(Z)] = 0$. However, the instrument relevance condition requires that $\mathrm{E}[g(Z)D] \neq 0$. $\mathrm{E}[g(Z)D]$ is also tightly tied to the efficiency of the IV estimator. Under homoskedasticity, the choice of $g(Z) = \eta_0(Z) = \mathrm{E}[D|Z]$ produces an asymptotically efficient estimator of $\theta_0$.

Importantly, only the AIPW score—also referred to as the "doubly robust" score—satisfies the key "Neyman orthogonality" property that is fundamental for DML. The IPW score is not Neyman orthogonal and should not be used together with generic machine learners. We verify Neyman orthogonality of the AIPW score in Appendix A, and illustrate empirical consequences of (non-)orthogonality in Section 4.2. □

Identification of the target parameter in each of these examples depends on nuisance parameters. Outside of special cases such as randomized controlled trials where the propensity score $\mathrm{E}[D|X] = r_0(X)$ is known by design, these nuisance parameters are generally unknown and thus need to be estimated.

In general, many moment conditions will exist for any target parameter. Examples 2 and 4 illustrate this by presenting two different moments that each identify the parameter of interest and could, in principle, be used for estimation. However, in both cases, only one of the proposed moment functions satisfies a key condition—Neyman orthogonality—that is crucial for obtaining reliable estimates in the presence of nuisance parameters.

In the next subsection, we discuss statistical issues stemming from the estimation of these nuisance parameters. Then, in Section 2.3, we outline how the combination of DML's two essential components alleviates the impact of nuisance estimation on the main inferential target. Fundamentally, it is this reduction of impact that allows DML to accommodate complex estimators, including ML methods, for nuisance estimation.

## 2.2 Impact of Nuisance Parameter Estimation

Suppose that we have at our disposal a first-step estimator, $\widehat{\eta}$, for the nuisance parameter $\eta_0$. This might be a parametric estimator such linear regression or a flexible, nonparametric learner from the ML toolbox. A plug-in estimator for the parameter of interest can be constructed as the solution to the sample average of the scores:

$$\widehat{\theta}: \quad \frac{1}{n}\sum_{i=1}^{n} m(W_i; \widehat{\theta}, \widehat{\eta}) = 0. \tag{13}$$

Inference about $\theta_0$ follows from a standard asymptotic Taylor expansion of (13) around the true parameters $(\theta_0, \eta_0)$:[8]

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} m(W_i;\widehat{\theta},\widehat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} m(W_i;\theta_0,\eta_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}m(W_i;\theta_0,\eta_0)(\widehat{\theta}-\theta_0) \\ + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}m(W_i;\theta_0,\eta_0)(\widehat{\eta}-\eta_0) + \text{higher order terms}\end{aligned}$$

$$\begin{aligned}\Rightarrow \quad \sqrt{n}(\widehat{\theta}-\theta_0) = -\Bigg( \Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}m(W_i;\theta_0,\eta_0)\Bigg]^{-1} \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(W_i;\theta_0,\eta_0)}_{\text{CLT}} \\ + \underbrace{\Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}m(W_i;\theta_0,\eta_0)\Bigg]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}m(W_i;\theta_0,\eta_0)(\widehat{\eta}-\eta_0)}_{(\star)\text{: First-order impact of nuisance estimation}} \\ + \sqrt{n}\times(\text{higher order terms})\Bigg),\end{aligned} \tag{14}$$

where "higher order terms" capture the impact of squared estimation errors $(\widehat{\eta}-\eta_0)^2$ and other remainders from the linearization. Under standard regularity conditions, the term labeled CLT in (14) will be approximately normal by a central limit theorem. The focus of DML is addressing the term $(\star)$.

The term $(\star)$ captures the first-order impact of estimating the nuisance parameter $\eta_0$. Its presence suggests the asymptotic distribution of $\widehat{\theta}$ will generally depend on the asymptotic behavior of the nuisance estimator $\widehat{\eta}$. That is, estimation error in $\widehat{\eta}$ propagates directly into inference about $\theta_0$, making the resulting distribution of $\widehat{\theta}$ differ from the idealized case in which $\eta_0$ is known. In situations where the nuisance parameter is low-dimensional (e.g., when assuming a linear model with few parameters), this additional uncertainty can be adequately characterized and managed through adjustments to the asymptotic variance; see, e.g., Section 6 of Newey and McFadden (1994).

However, the first-order dependence of $\widehat{\theta}$ on $\widehat{\eta}$ poses substantial complications in settings where the nuisance parameter is high-dimensional and $\widehat{\eta}$ corresponds to a flexible estimator. The complication results because flexible estimators are often associated with non-negligible bias and variance. As a consequence, (14) may be dominated by the first-order term $(\star)$ that involves $\widehat{\eta}$. In general, $(\star)$ *diverges* due to two issues, referred to as regularization bias and overfitting bias.

Regularization bias refers to the fact that neither term inside the sum in $(\star)$—

[8] We use a finite-dimensional expansion here to convey intuition. A formal treatment would require additional technicality to deal with cases where $\eta$ is a high- or infinite-dimensional object such as a function.

$\frac{\partial}{\partial \eta} m(W_i; \theta_0, \eta_0)$ and $(\widehat{\eta} - \eta_0)$—is mean zero in general. As a result, $(\star)$ is $\sqrt{n}$ times a sample average of a non-mean zero quantity, which does not converge in general. The term "regularization bias" reflects the fact that high-dimensional or nonparametric methods used to estimate $\eta_0$ often rely on regularization. That is, they introduce bias in order to control variance, implying that $(\widehat{\eta} - \eta_0)$ will not generally be mean zero in finite samples. Importantly, researchers have some control over score functions. The first key ingredient of DML—Neyman orthogonality, discussed in more detail in Section 2.3—is exactly the requirement that estimation be based on scores for which the high-dimensional analog of $\frac{\partial}{\partial \eta} m(W_i; \theta_0, \eta_0)$ is mean zero.

Overfitting bias, also referred to as own-observation bias, arises more subtly. Because $\widehat{\eta}$ is a function of the data used in its estimation, $\widehat{\eta} - \eta_0$ generally depends on the observations $W_i$ that are also used to construct the sample moment condition. This dependence typically occurs when the same dataset is used both to obtain $\widehat{\eta}$ and to evaluate the sample analog of equation (1), though it can arise more generally in dependent data settings. As a result, the product $\frac{\partial}{\partial \eta} m(W_i; \theta_0, \eta_0)(\widehat{\eta} - \eta_0)$ in $(\star)$ will generally not be mean zero, even if either term in the product is mean zero when considered in isolation.

We use the term "overfitting bias" to describe failures in conventional inference about the target parameter that arise from statistical dependence between $(\widehat{\eta} - \eta_0)$ and the observations used in the sample moment condition. We recognize that "overfitting" is typically associated with inflated variance in prediction contexts. Our usage emphasizes that overfitting in nuisance estimation can lead to bias in target parameter estimation, because it inflates the dependence between nuisance estimation error and the data.

The second key ingredient of DML—cross-fitting, also discussed in Section 2.3—addresses overfitting bias by using sample splitting to ensure (approximate) independence between the estimation error in the nuisance function and the observations used in the sample moment condition. This independence implies that the product in the numerator of $(\star)$ is mean zero whenever either term is mean zero.

The two main ingredients of DML, discussed in Section 2.3, are meant to alleviate the first-order impact of nuisance estimation. Estimation of nuisance parameters also generally affects the target parameter through the "higher order terms" in (14). A sufficient condition for these terms to be asymptotically ignorable is that the nuisance parameters are estimated accurately enough such that $\sqrt{n}\|\widehat{\eta} - \eta_0\|^2 \to_p 0$ under a suitable norm. A mean-square convergence rate faster than $n^{-1/4}$ is a commonly cited sufficient benchmark.[9]

Establishing such estimation quality guarantees for modern machine learning methods is an active area of research. Available results involve a combination of assumptions on

[9]Because the treatment of higher-order terms in DML is similar to that in other semiparametric approaches, we do not discuss them further. For more detailed discussion, see, e.g., Chernozhukov et al. (2018), Chernozhukov et al. (2022), and Kennedy (2023a).

the structure of the underlying data generating process and a choice of estimator that successfully leverages that structure. A canonical example is the lasso, which achieves suitable rates when the true regression function is sparse (e.g., Bickel, Ritov, and Tsybakov, 2009; Belloni et al., 2012). Related results for other classes of learners similarly require strong restrictions, such as smoothness, low effective dimension, or specific forms of regularization; see, for example, results for neural networks under compositional or smoothness assumptions (e.g., Farrell, Liang, and Misra, 2021b; Schmidt-Hieber, 2020) and for random forests under honesty and regularity conditions (e.g., Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019).

At the same time, recent work highlights important limitations of these results. In particular, many theoretical guarantees do not apply to off-the-shelf implementations with default tuning choices. For tree-based methods, available analyses show that, absent specific tree depth and subsampling choices, pointwise polynomial convergence rates may fail or that estimators may even be pointwise inconsistent, with only slow $L^2$ consistency established in high-dimensional settings (e.g., Chi et al., 2022; Cattaneo, Klusowski, and Yu, 2025). More broadly, if the nuisance function $\eta_0$ is fully nonparametric and high-dimensional without additional simplifying structure, no known method is able to attain the $n^{-1/4}$ rate required for standard DML asymptotics.

In applications, it is often unclear which assumptions are credible and which estimator is appropriate, making it difficult to judge whether the required convergence conditions plausibly hold. We thus return to the topic of the choice of ML estimator in Section 6.

## 2.3 Ingredients of DML

To reduce the dependence of estimators for $\theta_0$ on the estimation of high-dimensional nuisances $\eta_0$, DML estimators rely on two essential components—estimation based on Neyman orthogonal scores and cross-fitting—that mitigate regularization and overfitting bias. We now discuss each in turn.

### 2.3.1 Neyman Orthogonality

The chief difficulty in using the plug-in estimator $\widehat{\theta}$ is its dependence on the nuisance parameter estimator $\widehat{\eta}$. *Neyman orthogonality* is a local robustness property of the score function in (2) that decreases sensitivity of $\widehat{\theta}$ to errors in estimating $\eta_0$.

Formally, in addition to identifying the target parameter by satisfying (2), a *Neyman orthogonal score* also satisfies

$$\frac{\partial}{\partial \lambda}\Big\{\mathrm{E}\left[m(W;\theta_0,\eta_0+\lambda(\eta-\eta_0))\right]\Big\}\Big|_{\lambda=0}=0, \quad \forall \eta \in \mathcal{T}, \tag{15}$$

where $\eta$ denotes a candidate value for the nuisance parameter and $\lambda$ indexes the size of a local deviation away from the true $\eta_0$. Intuitively, (15) requires that, when we make a small move away from $\eta_0$ in any direction $\eta - \eta_0$, the moment condition remains unchanged.[10] In other words, small perturbations of the nuisance parameter away from the true value do not create first-order changes in the moment function. As a result, the plug-in estimator of $\theta_0$ is less sensitive to estimation error in $\widehat{\eta}$, thereby reducing regularization bias.

Returning to the expansion in (14), consider the case where the estimator $\widehat{\theta}$ is based on a Neyman orthogonal score. If the first-step estimation error $(\widehat{\eta} - \eta_0)$ were independent of the sample $\{W_i\}_{i=1}^n$ used to construct the score, Neyman orthogonality would imply that the second term in (14) vanishes,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}m(W_i;\theta_0,\eta_0)(\widehat{\eta}-\eta_0)\right)\approx 0, \tag{16}$$

under the convergence requirements on $\widehat{\eta}$ cited above and additional mild regularity conditions. Loosely, condition (15) means that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\eta}m(W_i;\theta_0,\eta_0)$ behaves like a mean 0 normalized sum, which converges by a central limit theorem. As long as $(\widehat{\eta}-\eta_0)$ converges to 0, the whole term disappears. Maintaining the assumption that higher order terms vanish, it follows that the asymptotic distribution of the plug-in estimator $\widehat{\theta}$ does not depend on the nuisance estimator $\widehat{\eta}$. Thus, inference about $\theta_0$ using $\widehat{\theta}$ can proceed as if $\eta_0$ were known. This heuristic derivation can be formalized and underlies the crucial importance of Neyman orthogonality in settings with complex nuisance parameters.

An important practical implication is that estimators based on Neyman orthogonal scores yield inference about $\theta_0$ that does not depend on the detailed statistical properties of the nuisance estimator $\widehat{\eta}$. This robustness is useful even in classical low-dimensional settings, where it avoids cumbersome variance adjustments or computationally costly resampling approaches to account for estimation of $\widehat{\eta}$. It becomes particularly important when modern flexible methods are used, as the statistical properties of these methods are still under active development. For example, only coarse rates of convergence are currently available for many promising machine learning methods. By alleviating the first-order impact of nuisance estimation, Neyman orthogonality makes it possible to combine such methods with standard asymptotic approximations, enabling formally valid inference for $\theta_0$ while exploiting modern flexible estimators for nuisance parameters.

We illustrate the importance of using Neyman orthogonal scores for obtaining reliable finite-sample inference in Section 4. There, we present simulation results demonstrating that estimators of the average treatment effect (ATE) based on non-orthogonal scores

[10] This formulation accommodates cases where $\eta_0$ is not finite-dimensional but instead a function or other complex object belonging to an abstract space $\mathcal{T}$.

may exhibit substantial bias, and the associated confidence intervals are often severely distorted. In contrast, DML estimators, which incorporate Neyman orthogonal scores as a core component, are approximately unbiased, and their confidence intervals achieve coverage rates close to the nominal level.

**Neyman orthogonal scores for common parameters.** Neyman orthogonal scores are well-known and readily available for common target parameters. We present Neyman orthogonal scores for six illustrative targets in Table 1. In Panels (i)-(iv), we present Neyman orthogonal scores corresponding to the target parameters in Examples 1-4 from Section 2.1. In Panel (v), we present an orthogonal score for the coefficient $\theta_0$ in the partially linear IV model

$$Y = \theta_0 D + g_0(X) + \varepsilon, \quad \mathrm{E}[Z\varepsilon] = \mathrm{E}[\varepsilon|X] = 0,$$

where $Z$ is an excluded scalar instrumental variable. This model differs from the simple IV model considered in Example 2 in two key respects: We allow for the presence of controls, and we do not impose full mean independence of the instrument from structural unobservables. Finally, in Panel (vi), we consider a more exotic target parameter—an average derivative corresponding to a continuous variable of interest. We present this case both because continuous variables are practically relevant and, more importantly, to highlight that nuisance parameters are not always conditional expectation functions (or projection coefficients as in the linear regression example).

**Deriving and verifying Neyman orthogonal scores.** Given the importance of Neyman orthogonal scores and their relevance for DML, we provide an outline of a general structure for obtaining Neyman orthogonal scores in Appendix B.

For a given score, one can generally verify Neyman orthogonality, or a lack thereof, by direct application of its definition in (15). We illustrate such a derivation for partially linear regression (Example 2) below. Similar derivations for the scores presented in Examples 1, 3, and 4 are provided in Appendix A.

**Example 2 (continued). Neyman Orthogonality of Partially Linear Regression Scores.** We introduced two score functions for identifying the partially linear regression coefficient $\theta_0$, (7) and (8). The first does not satisfy Neyman orthogonality, while the second—which corresponds to flexible partialling out—does.

Consider (7), which has nuisance parameter $g(X)$. Let $\Delta g(X) = g(X) - g_0(X)$; then

$$\frac{\partial}{\partial \lambda} \mathrm{E}[m_{naive}(W; \theta_0, g_0(X) + \lambda \Delta g(X))]\Big|_{\lambda=0} = \mathrm{E}\left[-\Delta g(X) D\right],$$

| *Observed Variables* ($W$) | *Nuisance parameters* ($\eta$) | *Neyman orthogonal score* ($m(W;\theta,\eta)$) |
|---|---|---|
| (i) *Target Parameter: Linear regression coefficient* | | |
| $Y$: outcome<br>$D$: treatment<br>$X$: controls | $\eta_{Y,0} = \arg\min_\eta \mathrm{E}[(Y - X'\eta)^2]$<br>$\eta_{D,0} = \arg\min_\eta \mathrm{E}[(D - X'\eta)^2]$ | $[(Y - X'\eta_Y) - \theta(D - X'\eta_D)]\,(D - X'\eta_D)$ |
| (ii) *Target Parameter: Partially linear regression coefficient* | | |
| $Y$: outcome<br>$D$: treatment<br>$X$: controls | $\ell_0(X) = \mathrm{E}[Y\|X]$<br>$r_0(X) = \mathrm{E}[D\|X]$ | $[(Y - \ell(X)) - \theta(D - r(X))]\,(D - r(X))$ |
| (iii) *Target Parameter: Linear IV coefficient (No controls)* | | |
| $Y$: outcome<br>$D$: treatment<br>$Z$: instruments | $r_0(Z) = \mathrm{E}[D\|Z]$ | $(Y - \theta D)\, r(Z)$ |
| (iv) *Target Parameter: Average treatment effect* ($\mathrm{E}\left[\mathrm{E}[Y\|D=1,X] - \mathrm{E}[Y\|D=0,X]\right]$) | | |
| $Y$: outcome<br>$D$: binary treatment<br>$X$: controls | $\ell_0(D,X) = \mathrm{E}[Y \mid D, X]$<br>$\alpha_0(D,X) = \frac{D}{r_0(X)} - \frac{1-D}{1-r_0(X)}$<br>for $r_0(X) = \mathrm{E}[D \mid X]$ | $\alpha(D,X)(Y - \ell(D,X)) + \ell(1,X) - \ell(0,X) - \theta$ |
| (v) *Target Parameter: Partially linear regression coefficient with excluded instruments* | | |
| $Y$: outcome<br>$D$: treatment<br>$Z$: instrument<br>$X$: controls | $\ell_0(X) = E[Y\|X]$<br>$r_0(X) = \mathrm{E}[D\|X]$<br>$h_0(X) = \mathrm{E}[Z\|X]$ | $[(Y - \ell(X)) - \theta(D - r(X))]\,(Z - h(X))$ |
| (vi) *Target Parameter: Average structural derivative* $\left(\mathrm{E}\left[\frac{\partial}{\partial d}\mathrm{E}[Y\|D=d,X]\,\|_{d=D}\right]\right)$ | | |
| $Y$: outcome<br>$D$: continuous treat.<br>$X$: controls | $\ell_0(D,X) = \mathrm{E}[Y\|D,X]$<br>$\alpha_0(D,X) =$<br>$-\frac{\partial}{\partial d}\log f_0(d,X)\,\|_{d=D}$ | $\frac{\partial}{\partial d}\ell_0(d,X)\,\|_{d=D} + \alpha(D,X)(Y - \ell(D,X)) - \theta$ |

*Notes:* For each target parameter, the table lists the observed variables, nuisance parameters, and a Neyman orthogonal score function that identifies the parameter. All expectations are taken conditional on the relevant covariates (e.g., $X$ or $Z$). In panel (vi), $f_0(D,X)$ denotes the conditional density of $D$ given $X$.

Table 1: Overview of common target parameters and Neyman orthogonal scores

which is generally non-zero when $D$ and $X$ are related. Hence, the score is not orthogonal.

Turning to (8), we have nuisance parameters $\eta(X) = (\ell(X), r(X))$ with true value $\eta_0(X) = (\ell_0(X) = \mathrm{E}[Y|X], r_0(X) = \mathrm{E}[D|X])$. Letting $\Delta\eta(X) = \eta(X) - \eta_0(X) = (\Delta\ell(X) = \ell(X) - \ell_0(X), \Delta r(X) = r(X) - r_0(X))$, we have

$$\begin{aligned}&\frac{\partial}{\partial\lambda}\mathrm{E}[m_{PLM}(W;\theta_0,\eta_0(X) + \lambda\Delta\eta(X))]\Big|_{\lambda=0}\\ &\quad = \mathrm{E}\left[-\Delta\ell(X)(D - r_0(X)) - \Delta r(X)(Y - \ell_0(X)) + 2\theta_0\Delta r(X)(D - r_0(X))\right] = 0\end{aligned}$$

where the last equality follows from $\ell_0(X) = \mathrm{E}[Y|X]$ and $r_0(X) = \mathrm{E}[D|X]$. □

The intuition for Neyman orthogonality in Example 2 is instructive and corresponds to

common intuition provided for partialling out. The orthogonal score uses only variation in $D$ and $Y$ that is (mean) independent of $X$, thereby isolating the identifying variation. As a result, small errors in one nuisance function can be offset by the other, making estimation more robust. In contrast, the non-orthogonal score uses all the variation in $D$ but adjusts only for the effect of controls on $Y$. Any mistakes in estimating $g(X)$ that are correlated with $D$ then directly bias the estimate of $\theta_0$, much like in a classic omitted variable scenario. Appendix B shows that the same partialling out intuition generalizes to average treatment effects and many other canonical target parameters.

### 2.3.2 Cross-fitting

Overfitting bias arises due to statistical dependence between the error in the nuisance parameter estimator and the data used in constructing the plug-in estimator. As discussed above, ignoring the first-step estimation of $\widehat{\eta}$ is justified only if the term $(\star)$ in (14) is asymptotically negligible. In the previous section, we outlined an argument for this term vanishing that depends on Neyman orthogonality *and* independence between the estimation error in the nuisance function, $\widehat{\eta} - \eta_0$, and the data used to construct the sample moment condition $\{W_i\}_{i=1}^n$. If $\widehat{\eta}$ is constructed using these same observations, the independence assumption will be violated. More generally, Neyman orthogonality alone is not sufficient to guarantee that first-step estimation of nuisance parameters can be ignored in inference about low-dimensional target parameters. To address this issue, DML relies on a second key ingredient: cross-fitting.

Cross-fitting is a form of repeated sample splitting intended to reduce the dependence between first-step estimation error and the data used to estimate the target parameter. Intuitively, if two independent datasets were available, one could be used to estimate the nuisance function $\widehat{\eta}$ and the other to estimate $\theta_0$ by plugging in $\widehat{\eta}$. In that case, independence between $\widehat{\eta} - \eta_0$ and $\frac{\partial}{\partial \eta} m(W_i; \theta_0, \eta_0)$ would follow by construction. With a single sample of independent observations, we can mimic this logic by randomly splitting the sample into two partitions, or "folds," using one fold to estimate $\widehat{\eta}$ and the other to evaluate the score function and estimate $\theta_0$. Of course, such an approach inefficiently uses the available data because both $\eta_0$ and $\theta_0$ are estimated using only subsets of the data. Cross-fitting restores asymptotic efficiency by rotating which folds are used for each task, so all observations contribute to both steps.

The cross-fit version of the plug-in estimator with a generic score defined in (13) is

$$\widehat{\theta}^{\,\times\text{-fit}}: \quad \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} m(W_i; \widehat{\theta}^{\,\times\text{-fit}}, \widehat{\eta}_{-k}) = 0. \tag{17}$$

Here, $\{I_k\}_{k=1}^K$ is a random partition of the sample of units $\{1, \ldots, n\}$ into $K$ subsamples

of approximately equal size, and $\widehat{\eta}_{-k}$ denotes a first-step nuisance parameter estimate constructed using only observations *excluding* those in the subsample $I_k$. Because $\widehat{\eta}_{-k}$ uses only observations *not* in subsample $k$, the estimation error in $\widehat{\eta}_{-k}$ is independent of observations in subsample $k$, which alleviates overfitting bias. By rotating ("crossing") samples for the estimation of nuisance parameters and the estimation of target parameters, cross-fitting asymptotically avoids losing efficiency relative to the hypothetical full-sample estimator that makes use of the true values of the nuisance parameters.

As a byproduct that comes at no additional computational cost, cross-fitting produces out-of-sample prediction errors associated with the nuisance functions. These enable calculation of diagnostics for evaluating the choice and specification of the nuisance function estimator. For example, in partially linear regression (Example 2), we have $\eta_0 = (\ell_0, r_0)$ with $\ell_0(X) = E[Y|X]$ and $r_0(X) = E[D|X]$. The *cross-fitted* errors $(Y_i - \widehat{\ell}_{-k}(X_i))$ and $(D_i - \widehat{r}_{-k}(X_i))$—where $\widehat{\ell}_{-k}$ and $\widehat{r}_{-k}$ are estimated using data not containing observation $i$—are equivalent to *cross-validated* errors from $K$-fold cross-validation. Their availability allows calculating metrics for learner performance such as mean-squared prediction error (MSPE) and out-of-sample $R^2$. We discuss these and other methods for evaluating the nuisance function estimators in Section 6.

The simulation results in Section 4 provide finite-sample evidence that cross-fitting plays an important role in delivering reliable inference. The results illustrate that cross-fitting while using non-orthogonal scores delivers relatively little benefit in the considered examples. Similarly, the results show that inference based on Neyman orthogonal scores *without* cross-fitting often fails to deliver reliable inferential results. In contrast, DML estimators, which incorporate both Neyman orthogonal scores and cross-fitting, deliver the most robust performance across the considered examples.

**Remark 1** (Cross-fitting with Dependence)
Cross-fitting based on random partitions of $\{i : 1, \ldots, n\}$ applies to cross-sectional and fixed-$T$ panel settings with independence across $i$ and arbitrary temporal dependence. In panel settings, the sample is partitioned by cross-sectional units, preserving the full time series per unit. Cross-fitting can also be extended to more complex dependence structures. For example, Chiang et al. (2022) extend DML to settings with multiway clustered dependence, and Semenova et al. (2023) and Ballinari and Wehrli (2025) discuss its application in time series and dynamic panels with weak dependence.

**Remark 2** (Alternatives to Cross-fitting)
Under special conditions on the structure of the data, overfitting can be avoided by

carefully tailoring nuisance estimators, bypassing the need for cross-fitting. This approach is taken in, for example, Belloni et al. (2012), Belloni, Chernozhukov, and Hansen (2014), van de Geer et al. (2014), Javanmard and Montanari (2014), Zhang and Zhang (2014), Belloni et al. (2017), Farrell, Liang, and Misra (2021b), and Wiemann (2026). Such results are available for relatively few ML tools. Simulations also suggest that DML with flexible learners can perform similarly to these procedures when their conditions hold and outperforms them otherwise (e.g., Ahrens et al., 2025).

# 3 Estimation and Inference with DML

In this section, we define the DML estimator, present an implementation algorithm, and summarize its asymptotic properties. We remain in the general semiparametric setting of Section 2. That is, our focus is estimation and inference for a low-dimensional target parameter $\theta_0$, which is defined by moment conditions (2) and depends on an unknown (potentially high-dimensional) nuisance parameter $\eta_0$.

Within this general setting, a DML estimator is a plug-in estimator that combines *both* Neyman orthogonal scores and cross-fitting. The DML estimator of $\theta_0$ is then

$$\widehat{\theta}_{DML}: \quad \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} m(W_i;\widehat{\theta}_{DML},\widehat{\eta}_{-k}) = 0, \tag{18}$$

where $m$ is a Neyman orthogonal score (i.e., a score that satisfies (2) and (15)), $\{I_k\}_{k=1}^K$ is a random partition of the sample of individuals $\{1,\ldots,n\}$ into $K$ subsamples of approximately equal size, and $\{\widehat{\eta}_{-k}\}_{k=1}^K$ are cross-fitted nuisance parameter estimators.

Chernozhukov et al. (2018) provide conditions that allow researchers to make valid inferential statements about $\theta_0$ using the DML estimator $\widehat{\theta}_{DML}$. These conditions involve conventional sampling and regularity conditions along with the assumption that the nuisance parameter estimator, $\widehat{\eta}$, converges sufficiently quickly, as discussed in Section 2.2. Under these assumptions, the DML estimator is asymptotically normal:

$$\sqrt{n}\widehat{\Sigma}^{-1/2}(\widehat{\theta}_{DML}-\theta_0) \xrightarrow{d} \mathcal{N}(0,\mathrm{I}),$$

where

$$\begin{aligned}\widehat{\Sigma} =& \widehat{J}^{-1}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} m(W_i;\widehat{\theta}_{DML},\widehat{\eta}_{-k})m(W_i;\widehat{\theta}_{DML},\widehat{\eta}_{-k})^\top\right)\widehat{J}^{-1\top},\\ \widehat{J} =& \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{\partial}{\partial\theta}\left\{m(W_i;\theta,\widehat{\eta}_{-k})\right\}\Big|_{\theta=\widehat{\theta}_{DML}}.\end{aligned} \tag{19}$$

Standard errors for $\widehat{\theta}$ are thus given by square-roots of the diagonal values of $\widehat{\Sigma}/n$.

A key practical implication of using both Neyman orthogonal scores and cross-fitting is that standard errors for $\widehat{\theta}_{DML}$ can be computed as if the nuisance functions were known. This approximation result holds in a variety of settings, including standard cross-sectional data, clustered cross-sectional data, and panel data with fixed $T$. In cross-sectional applications, the standard errors coincide with conventional heteroskedasticity-robust formulas. In clustered or panel settings, they are analogous to clustered standard errors that allow for arbitrary dependence across time within units $i$.

**Remark 3** (Semiparametric Efficiency of DML Estimators)
The DML estimator $\widehat{\theta}_{DML}$ is semiparametrically efficient when the orthogonal score coincides with the efficient influence function of the target parameter. Methods for deriving such scores are well-established; see, for example, Newey (1994) and Chernozhukov et al. (2022).

As examples, the orthogonal scores in Examples 1-3 correspond to the efficient influence functions for their respective target parameters under homoskedasticity. The scores for the ATE in Example 4 and for the group-time average treatment effect on the treated in Section 5 are likewise efficient for their parameters. Thus, the resulting estimators are semiparametrically efficient.

Algorithm 1 illustrates computation of $\widehat{\theta}_{DML}$ in an *i.i.d.* setting. Implementation proceeds in three parts. First, data are randomly split into $K$ subsamples (Step 1). Second, the cross-fitted nuisance estimates are computed in each subsample (Steps 2–4). Third, estimation and inference about the target parameter takes place (Steps 5–6).

**Algorithm 1** DML Estimation and Inference

**Require:** A sample $\{W_i\}$ for $i \in \{1, \ldots, n\}$, a Neyman orthogonal score $m$, a nuisance parameter estimator $\widehat{\eta}$, an integer $K$ for the number of cross-fitting folds.

1: Randomly split the sample of indices $\{1, \ldots, n\}$ into $K$ partitions $(I_k)_{k=1}^K$ of approximately equal size.
2: **for** k = 1 to $K$ **do**
3: Compute the nuisance parameter estimator on samples $I_k^c = \{1, \ldots, n\} \setminus I_k$:

$$\widehat{\eta}_{-k} = \widehat{\eta}(\{W_i\}_{i \in I_k^c}).$$

4: **end for**
5: Construct the DML estimator $\widehat{\theta}_{DML}$ for $\theta_0$ as the solution to (18).
6: Estimate the covariance matrix $\widehat{\Sigma}$ via (19).

Importantly, the algorithm is highly general: it is a blueprint for estimating and performing inference on a broad range of target parameters $\theta_0$. The procedure accommodates a wide range of nuisance parameter estimators $\widehat{\eta}$, including classical approaches as well as modern and emerging ML methods.

Implementing Algorithm 1 requires researchers to make several design choices. These include selecting an appropriate scheme for generating the cross-fitting folds, choosing the number of cross-fitting folds, and specifying the nuisance function estimator. We discuss these and other implementation choices in the following sections. In Section 7, we provide a discussion of current best practices for the implementation of DML estimators.

# 4 Two Simulation Illustrations

This section employs two simulation exercises to illustrate the consequences of relying on non-Neyman orthogonal scores and of failing to perform cross-fitting when using flexible nuisance estimators such as ML methods. We discuss two simulations to underscore that bias dominates in some contexts, while regularization bias prevails in others.

The first simulation focuses on the role of cross-fitting in the linear IV setting with many instruments. This problem has been thoroughly studied since the mid-1990s (e.g., Bekker, 1994). One of the recommendations to emerge from this literature is to use sample-splitting to reduce overfitting bias (often termed "many instruments bias" in this context). Our simulation revisits this classical many instrument setup by comparing linear IV estimation with ML-based IV estimation. As each of these approaches relies on Neyman orthogonal scores (see Example 3), the IV setting isolates the impact of overfitting. We show that cross-fitting alleviates overfitting bias and that using first stage estimators other than OLS can lead to efficiency gains.

Our second example turns to the estimation of average treatment effects, where we emphasize the role of Neyman orthogonality in delivering valid inference. Using a calibrated simulation, we compare estimators based on orthogonal and non-orthogonal scores, with and without cross-fitting. The results highlight that only the combination of orthogonality and cross-fitting yields estimators with reliable sampling behavior, thereby illustrating why these two ingredients are central to the DML framework.

## 4.1 Instrumental Variables with Many Instruments

We present simulation results from the canonical many-instrument linear IV model (see Example 3). We generate data as *i.i.d.* realizations from

$$Y = \theta_0 D + \varepsilon, \qquad D = g_0(Z) + \nu, \qquad (\varepsilon, \nu)' \perp Z.$$

Here $Z$ are $p = 200$ simulated instruments, $\theta_0 = 0$ is the parameter of interest, and $(\varepsilon, \nu)'$ are correlated error terms drawn from the normal distribution. We define the nuisance function as $g_0(Z) = Z'\pi_0$, where the first six elements of $\pi_0$ are set to 0.1 and the remaining entries are set to 0. That is, only the first six instruments carry any signal. In our setting, OLS suffers from an upward bias of around 0.5. All results are based on a sample size of $n = 1000$ and 1000 simulation replications.[11]

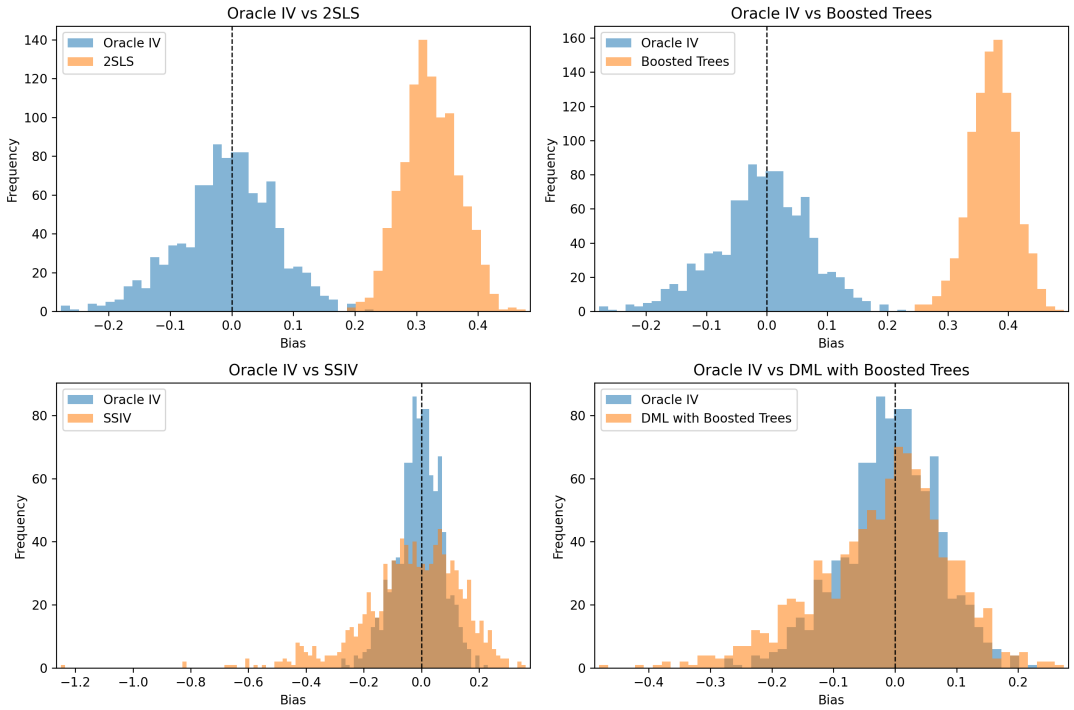
We compare five estimators, each based on a Neyman orthogonal score. As a benchmark, we report results from the oracle IV estimator that knows the true first-stage function and uses $Z'\pi_0$ as the instrument (Oracle IV). This estimator is infeasible in practice but serves as an efficient baseline. We next consider two-stage least squares using all 200 instruments (2SLS). As is well-known from the many-instruments literature, 2SLS performs poorly when the number of instruments is large. We implement DML with a linear first stage regression using all 200 instruments. Note that this DML estimator is essentially split-sample IV (SSIV; Angrist and Krueger, 1995a; Angrist, Imbens, and Krueger, 1999).[12] One can thus view DML in this context as an extension of SSIV that also accommodates the use of flexible first stage estimators, for example, because the researcher does not want assume linearity and is unsure about which instruments matter. To this end, we also estimate the first stage using gradient boosted trees, both without cross-fitting (Boosted Trees) and with cross-fitting (DML with Boosted Trees). While one should consider alternative learner choices in practice, we focus on gradient boosted trees with a maximum tree depth of four, a learning rate of 0.1, and early stopping based on a 20% validation set. We use 5-fold cross-fitting for SSIV and DML with gradient boosted trees.

We report results in Figure 1 and Table 2. As expected and predicted by the theory, the methods that do not use cross-fitting—2SLS and Boosted Trees—perform poorly. Although the distributions of estimates are tightly concentrated, they are centered far

[11] Specifically, we generate $Z \sim N(0_p, .4\mathrm{I}_p + .6\iota_p\iota_p')$ and $(\varepsilon, \nu)' = N(0_2, \Sigma)$ with $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{21} = \Sigma_{21} = .6$. We use $0_p$ and $\iota_p$ to denote a $p \times 1$ vector of zeros and ones, respectively; $\mathrm{I}_p$ denotes a $p \times p$ identity matrix.

[12] More precisely, Angrist and Krueger (1995a) suggest splitting the sample randomly into two folds, using the first fold to estimate the IV first stage via linear regression, and the second to estimate the target parameter. That is, they do not cross-fit, though note the possibility of doing so in their conclusion. Furthermore, DML with a linear (first-stage) regression and $K = n$ is exactly jackknife IV (Angrist, Imbens, and Krueger, 1999).

Figure 1: Histograms of IV Coefficient Estimates




*Notes.* Each panel compares simulation performance of a feasible IV estimator (orange) to the infeasible oracle IV estimator (blue).

Table 2: IV simulation results by estimation method

| Variable | Bias | Median Bias | Std. Dev. | Coverage |
|---|---|---|---|---|
| Oracle IV | -0.0076 | -0.0036 | 0.0773 | 0.9570 |
| 2SLS | 0.3228 | 0.3199 | 0.0447 | 0.0000 |
| Boosted Trees | 0.3738 | 0.3751 | 0.0370 | 0.0000 |
| SSIV | -0.0325 | -0.0172 | 0.1725 | 0.9380 |
| DML with Boosted Trees | -0.0201 | -0.0051 | 0.1075 | 0.9590 |

*Notes:* This table presents simulation summary statistics of estimators $\theta_0$ as described in the main text. Coverage denotes the simulation coverage of 95% confidence intervals based on homoskedastic standard errors. All estimators use Neyman orthogonal scores. SSIV and DML with Boosted Trees also use cross-fitting.

from the truth, yielding average biases an order of magnitude larger than estimators relying on cross-fitting. As a result, their 95% confidence intervals exhibit zero coverage across the 1000 simulation replications. In contrast, the estimators with cross-fitting, SSIV and DML with Boosted Trees, perform substantially better. They have relatively small bias and produce coverage near the nominal 95% level. Compared to the infeasible oracle IV, they exhibit a visually larger spread, but are also approximately centered around zero.

Importantly, we observe a meaningful difference between the two cross-fitted estimators. DML with Boosted Trees outperforms SSIV, reflecting the advantage of using a more flexible and regularized first-stage learner in this context. While this is specific to the design considered here, it highlights a broader point: the choice of learner can matter

substantially in practice. We return to this theme in the empirical example in Section 6.

Finally, we emphasize that all five estimators rely on Neyman orthogonal scores. As such, these simulations are designed to isolate the impact of cross-fitting in reducing the impact of overfitting during nuisance parameter estimation. In the next subsection, we shift focus to the role of orthogonality itself by comparing estimators with and without Neyman orthogonal scores in the context of average treatment effect estimation.

## 4.2 Average Treatment Effect Estimation

This section uses a calibrated simulation to illustrate the importance of Neyman orthogonality for valid inference. We focus on estimation of the average treatment effect (ATE) under unconfoundedness—that is, assuming treatment is as good as randomly assigned after conditioning on covariates. This setting is instructive as it allows comparison of DML to the IPW estimator, which is commonly used despite not being based on a Neyman orthogonal score.

The simulation is based on Poterba, Venti, and Wise (1995), who study the effect of 401(k) eligibility ($D$) on household net financial assets ($Y$), treating eligibility as random given observed covariates. The observed covariates ($X$), include age, income, education, family size, and indicators for two-earner households, home ownership, and alternative pension coverage. We use this application because it is a standard example in work on treatment effect estimation with machine learning; see, for instance, Belloni et al. (2017), Chernozhukov et al. (2018), Wüthrich and Zhu (2023), and Ahrens et al. (2025).

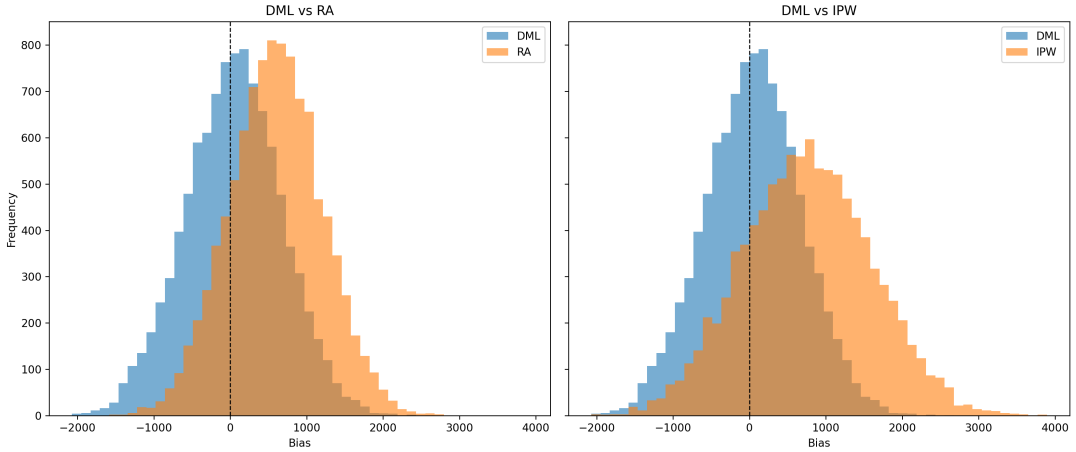
We base the simulation on the same 9915 observations used in the studies cited above. We calibrate the data-generating process for the simulation by flexibly estimating the propensity score and the conditional outcome model. We then use the covariate values for each individual in the original data to simulate new treatment assignments and outcomes. Under this design, the true value of the ATE is approximately 6,889.[13]

We compare three scores for estimating the ATE: the IPW score, the Neyman orthogonal AIPW score, and the regression adjustment (RA) score. The IPW and AIPW scores are defined in Example 4. The RA score is given by

$$m_{RA}(W;\theta,\eta) = \ell(1,X) - \ell(0,X) - \theta, \tag{20}$$

[13]Specifically, we estimate the probability of 401(k) eligibility given covariates, $r_0(X) = \mathrm{P}(D = 1 \mid X)$, and the conditional mean of financial assets given treatment and covariates, $\ell_0(D,X) = \mathrm{E}[Y \mid D, X]$, using random forests trained on the full sample (1,000 trees and minimum node size 10). We estimate treatment-specific residual variances, $\sigma_d^2$, as the sample variance of $y_i - \widehat{\ell}(d_i, x_i)$ within treatment groups. In each simulation replication, we draw $d_{i,s} \sim \text{Bernoulli}(\widehat{r}(x_i))$ and generate outcomes as $y_{i,s} = \widehat{\ell}(d_{i,s}, x_i) + \varepsilon_{i,s}$, where $\varepsilon_{i,s} \sim N(0, \widehat{\sigma}^2_{d_{i,s}})$.

Figure 2: Histograms of Cross-fit ATE Estimators




*Notes.* Each panel compares simulation performance of the DML estimator (blue) to a cross-fit ATE estimator based on a score that is not Neyman orthogonal (orange). Specifically, the left panel compares DML to a cross-fit regression adjustment (RA) estimator, and the right panel compares DML to a cross-fit IPW estimator.

where the nuisance parameter is $\eta(D, X) = \ell(D, X)$ with true value $\ell_0(D, X) = \mathrm{E}[Y|D, X]$. Like the IPW score, the RA score is not Neyman orthogonal. For each score, we report estimates both with and without cross-fitting. The DML estimator of the ATE corresponds to using the AIPW score with cross-fitting. All estimators use the same tuning parameter choices for nuisance function estimation, so differences in performance reflect only the choice of score and use of cross-fitting. We use random forests with 1000 trees and a maximum depth of 8 and 4 for the outcome and treatment propensity functions, respectively. We set the number of folds to $K = 10$.

Figure 2 is designed to highlight the role of Neyman orthogonality by comparing the DML estimator to two alternatives based on non-orthogonal scores: the cross-fit RA estimator (left panel) and the cross-fit IPW estimator (right panel). The most striking feature in both comparisons is that DML appears approximately unbiased, while the RA and IPW estimators exhibit substantial bias. The distribution of DML is also more concentrated than that of IPW and similar in spread to RA. Taken together, the plots show that DML dominates in this example due to its much smaller bias and comparable or better precision.

Table 3 reports simulation results across a range of performance metrics. The DML and AIPW estimators, which are each based on Neyman orthogonal scores, dominate the remaining procedures. They exhibit substantially lower bias and mean absolute deviation than the IPW and RA estimators, whether or not cross-fitting is used. While the RA estimators have a slightly smaller standard deviation than DML and AIPW, this is offset by their much larger bias, implying inferior performance for most reasonable loss functions.

DML and AIPW also exhibit superior coverage performance. For these two estimators, we report confidence intervals based on the standard error estimator defined in (19).

Table 3: ATE Simulation Results

| | $\widehat{\theta}_{DML}$ | $\widehat{\theta}_{RA}^{\times\text{-fit}}$ | $\widehat{\theta}_{IPW}^{\times\text{-fit}}$ | $\widehat{\theta}_{AIPW}$ | $\widehat{\theta}_{RA}$ | $\widehat{\theta}_{IPW}$ |
|---|---|---|---|---|---|---|
| Bias | 47.3 | 594.9 | 756.4 | 142.0 | 644.0 | 747.3 |
| Median Bias | 60.8 | 597.8 | 755.1 | 154.7 | 647.1 | 745.6 |
| Mean Abs. Dev. | 502.2 | 701.1 | 923.7 | 508.8 | 734.4 | 914.9 |
| Std. Dev. | 627.4 | 606.2 | 841.7 | 619.2 | 604.7 | 837.3 |
| Coverage | 0.945 | 0.834 | 0.851 | 0.911 | 0.815 | 0.853 |
| Neyman orthogonal | Yes | No | No | Yes | No | No |
| Cross-fitting | Yes | Yes | Yes | No | No | No |

*Notes:* $\widehat{\theta}_{AIPW}$, $\widehat{\theta}_{RA}$, and $\widehat{\theta}_{IPW}$ respectively denote estimation based on the AIPW, RA, and IPW scores without cross-fitting. Similarly, $\widehat{\theta}_{RA}^{\times\text{-fit}}$, and $\widehat{\theta}_{IPW}^{\times\text{-fit}}$ respectively denote estimation based on the RA and IPW scores with cross-fitting. $\widehat{\theta}$ is the DML estimator, which is based on the AIPW (Neyman orthogonal) score with cross-fitting. Coverage is simulation coverage of 95% confidence intervals. For both $\widehat{\theta}_{DML}$ and $\widehat{\theta}_{AIPW}$, intervals are computed using the standard error estimator implied by (19). For the remaining estimators, intervals are computed using the infeasible simulation standard deviation. Results are based on 10000 simulation replications.

This estimator is theoretically justified for DML under relatively weak conditions and would be valid for AIPW if overfitting were sufficiently controlled. In contrast, for the non-orthogonal IPW and RA estimators, we report coverage using the simulation standard deviation, since their first-order behavior depends directly on the nuisance function estimator, which makes valid standard error estimation challenging.

Even using the infeasible simulation standard deviation, the IPW and RA estimators substantially undercover the true ATE. While substantially better than IPW or RA, we see that the AIPW estimator also undercovers. In contrast, DML achieves near-nominal coverage using estimated standard errors, and outperforms AIPW in both bias and coverage.[14] This highlights the central insight of the simulation: valid inference relies on combining Neyman orthogonality with cross-fitting.

# 5 Economic Consequences of Hospital Admission

To demonstrate the flexibility of DML, we apply it in a staggered adoption panel setting to estimate group-time average treatment effects on the treated and dynamic average treatment effects as discussed in Callaway and Sant'Anna (2021). Outside of illustrating the application of DML in a canonical panel data setting, we further use this example to discuss the additional randomness introduced in DML due to the use of sample splitting. We show that simply repeating the DML estimation can help us gauge the robustness of conclusions to particular sample splits. We then outline the median aggregation approach presented in Chernozhukov et al. (2018) as a way to summarize results across repetitions.

Our example builds on Dobkin et al. (2018) and Sun and Abraham (2021). Dobkin

[14] If we used simulation standard deviations for DML and AIPW, coverage would be 0.948 and 0.945.

et al. (2018) analyze the causal effect of hospital admission on several economic outcomes, including out-of-pocket medical spending, using conventional two-way fixed effects applied to a panel of U.S. households from the Health and Retirement Study (HRS). Sun and Abraham (2021) extend this analysis by estimating dynamic effects using more flexible methods that allow for treatment effect heterogeneity. We use the same data as Sun and Abraham (2021), which consist of a balanced panel of 656 households observed over waves 7 through 11 of the HRS.[15]

We proceed in two steps. In Section 5.1, we estimate the change in out-of-pocket medical spending at time period $t$ caused by hospitalization at time $g$ for all potential pairs $(t, g)$ in the data. In Section 5.2, we then aggregate these estimates for inference about the dynamic effects of hospitalization.

## 5.1 Group-Time Average Treatment Effects of Hospitalization

The group-time average treatment effect on the treated (GT-ATT) measures the effect of hospitalization at time $g$ on medical spending at time $t \leq T$, for individuals first hospitalized in period $g$. While the GT-ATTs are often not of primary interest, they can be aggregated to population-weighted ATTs or dynamic effects. We follow the potential outcome setup of Callaway and Sant'Anna (2021) to formally define the GT-ATT. Let $G_i$ denote the time of first hospital admission for individual $i$. Let $Y_{i,t}(0)$ denote the potential outcome of out-of-pocket medical spending of individual $i$ at time $t$ if they remain untreated throughout, and let $(Y_{i,t}(g))_{g=2}^{T}$ be the set of potential outcomes at time $t$ for every potential time $g$ of the first hospital admission. Observed outcomes are given by $Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{T}(Y_{i,t}(g) - Y_{i,t}(0))\mathbb{1}\{G_i = g\}$. Then, the GT-ATT for group $g$ and time $t$ is defined as

$$\theta_0^{(g,t)} = \mathrm{E}[Y_{i,t}(g) - Y_{i,t}(0)|G_i = g]. \tag{21}$$

Following Dobkin et al. (2018) and Sun and Abraham (2021), we leverage a parallel trends assumption on not-yet-hospitalized individuals to identify the GT-ATTs. Motivated by the robustness checks of Dobkin et al. (2018), we consider a more flexible version of this assumption: parallel trends are assumed to hold only after conditioning on observed pre-treatment characteristics, such as age, gender, race, and education. Under this conditional parallel trends assumption and additional standard conditions (see, e.g., Callaway and Sant'Anna, 2021), the GT-ATT is equivalent to

$$\theta_0^{(g,t)} = \mathrm{E}[\Delta_g Y_{i,t}|G_i = g] - \mathrm{E}[\mathrm{E}[\Delta_g Y_{i,t}|G_i \neq g, G_i > t, X_i]|G_i = g],$$

[15] See Table 2 of Sun and Abraham (2021) for summary statistics of the HRS sample we use.

where $\Delta_g Y_{i,t} = Y_{i,t} - Y_{i,g-1}$ is the difference of outcomes in period $t$ and the baseline period $g-1$. A standard no anticipation assumption implies $\theta_0^{(g,t)} = 0$ for $t < g$, which is often used to test for pre-trends.

A key insight about the estimation of the GT-ATT is that for each group-time pair, the target parameter is equivalent to a conventional ATT identified under conditional unconfoundedness where the outcome is replaced by an appropriate difference of outcomes. Building on the efficient ATT score of Hahn (1998), we employ an augmented inverse probability weighted (AIPW) score for the group–time average treatment effect (GT-ATT) as the basis for DML:[16]

$$\begin{aligned} m^{(g,t)}(W_i; \theta, \eta) = & \frac{\mathbb{1}\{G_i = g\}\Big(\Delta_g Y_{i,t} - \ell^{(g,t)}(X_i)\Big)}{\pi^g} \\ & - \frac{q^{(g,t)}(X_i)\, \mathbb{1}\{G_i \neq g\}\mathbb{1}\{G_i > t\}\Big(\Delta_g Y_{i,t} - \ell^{(g,t)}(X_i)\Big)}{\pi^g\Big(1 - q^{(g,t)}(X_i)\Big)} \\ & - \frac{\mathbb{1}\{G_i = g\}}{\pi^g}\,\theta \end{aligned} \tag{22}$$

where $\mathbb{1}\{G_i = g\}$ denotes the treated group, $\mathbb{1}\{G_i \neq g\}\mathbb{1}\{G_i > t\}$ denotes units not yet treated by time $t$ that serve as controls, and the nuisance parameter $\eta = \Big(q^{(g,t)}, \ell^{(g,t)}, \pi^g\Big)$ takes true values $q_0^{(g,t)}(X_i) = \mathrm{P}\Big(G_i = g \mid X_i, \{G_i = g\} \cup \{G_i > t\}\Big)$, $\ell_0^{(g,t)}(X_i) = \mathrm{E}[\Delta_g Y_{i,t} \mid G_i \neq g,\ G_i > t,\ X_i]$, and $\pi_0^g = \mathrm{P}(G_i = g)$. One can confirm that the score (22) satisfies both the moment condition (2) and Neyman orthogonality (using steps similar to Example 4, Appendix A).

The application of DML to estimation of the GT-ATTs provides a useful robustness check because it seems unlikely that a researcher has prior knowledge of the parametric form of the nuisance functions. Even in this application with only few control variables, it is practically impossible to saturate the model, which would eliminate any additional need for flexible estimation. A fully interacted set of the eight control variables results in 3,072 indicators, which exceeds the number of households in the sample.

We set the number of cross-fitting folds to $K = 15$. Since the number of observations per group-time sample is relatively small, we opt for more folds than the rule of thumb of 5-10 folds often cited in the literature on cross-validation. For ease of exposition, we report only results on a single machine learner: a random forest estimator with 1000 trees and a minimum node size of 10 to estimate the nuisance parameters.

Table 4 presents DML estimates for all identified group-time average treatment effects. Columns (1) through (5) show results from five different random partitions of the data

[16] Our treatment mirrors Callaway and Sant'Anna (2021). The AIPW score directly builds on Chang (2020), who proposes a DML estimator in the canonical two-group, two-period difference-in-differences design. See also Sant'Anna and Zhao (2020).

used for cross-fitting. Although these DML estimators are asymptotically equivalent, they can yield different results in finite samples due to variation across sample splits. This variability may be particularly pronounced in settings with relatively few observations, as is often the case in staggered adoption designs.

To illustrate, consider the first-period treatment effect for group 9. This estimate is based on only 176 treated and 228 untreated observations. The corresponding DML estimates obtained from the five different considered sample splits range from 3204 to 3486—a difference of roughly 30% of the corresponding standard errors. While this variation is not large enough to change the sign or significance of the estimate, it shows that DML can produce meaningful differences in practice depending on the sample split.

We believe that replicating the DML procedure several times is important for understanding the variability induced by sample splitting. While under ideal conditions the DML estimator from any one split is asymptotically equivalent to that from any other, finite-sample differences may arise. More importantly, large variation across splits may signal deeper issues. For example, substantial across-split variation may signal that asymptotic approximations are unreliable in the sample at hand (e.g., due to heavy tails, poor overlap, or unstable nuisance estimates), or that the underlying machine learners are unstable within the available data.

However, reporting all results may lead to information overload or be practically infeasible. We therefore recommend that researchers examine the full set of estimates and standard errors across splits and ensure they are available for transparency. Following Chernozhukov et al. (2018), we suggest median aggregation as a simple and practical strategy for summarizing results.[17] Letting $\widehat{\theta}_s$ and $\widehat{s.e.}_s$ denote the $s^{\text{th}}$ DML estimator and its standard error across $S$ replications, the median-aggregated point estimate and standard error are

$$\begin{aligned} \widehat{\theta}^{\text{median}} &= \text{median}\left(\{\widehat{\theta}_s\}_{s=1}^{S}\right), \\ \widehat{s.e.}^{\text{median}} &= \sqrt{\text{median}\left(\{\widehat{s.e.}_s^2 + (\widehat{\theta}_s - \widehat{\theta}^{\text{median}})^2\}_{s=1}^{S}\right)}, \end{aligned} \tag{23}$$

where the median is applied element-wise for vector-valued parameters. The standard error accounts for both conventional sampling uncertainty and variability across splits.

Column (6) of Table 4 presents the median-aggregated estimators for the identified GT-ATTs. The results strengthen the economic conclusions of Dobkin et al. (2018) and Sun and Abraham (2021). When we use DML to estimate the same effects allowing for parallel trends to hold only conditional on controls and using a flexible method to include those controls, we continue to estimate that hospitalization causes a substantial and significant

[17] Chernozhukov et al. (2018) also consider mean aggregation. We are not aware of formal arguments favoring one approach over another in finite samples. Our recommendation is heuristic.

increase in out-of-pocket medical spending in the period immediately following the event. This conclusion holds for all three cohorts considered in our analysis.

We also compare the DML estimates to those obtained using the parametric AIPW estimator proposed by Sant'Anna and Zhao (2020). This estimator relies on the same AIPW score as DML but differs in how the nuisance functions are estimated. Specifically, it assumes that the propensity score $\mathrm{P}\Big(G_i = g \mid X_i, \{G_i = g\} \cup \{G_i > t\}\Big)$ follows a logit model with a linear index and that the outcome regression $\mathrm{E}[\Delta_g Y_{i,t} \mid G_i \neq g,\ G_i > t,\ X_i]$ is linear in $X_i$. If both parametric models are correctly specified, the estimator is first-order equivalent to DML. If either is misspecified, inference remains valid but efficiency may be lost. If both are misspecified, the estimator is inconsistent. Because these parametric forms are typically motivated by convenience rather than theory, DML offers an appealing alternative. In practice, the choice between the two depends on the researcher's confidence in the parametric models and access to a flexible nuisance estimator of reasonable quality. Regardless of the preferred approach, using the other as a robustness check is a sensible practice.

Columns (7) and (8) of Table 4 present the Sant'Anna and Zhao (2020) estimator without and with controls, respectively. The comparison confirms that adjusting for covariates materially affects the results. Relative to the size of the standard errors, the median-aggregated DML estimates (in Column 6) and the AIPW DiD estimates (Column 8) are qualitatively similar. The most notable exception is group 9's treatment effect in Wave 10. The parametric estimate is 2534 ($s.e. = 422$) and thus substantially larger than the DML estimate of 937 ($s.e. = 434$). Further, group 10's pre-treatment effect in Wave 7 is negative under DML ($-1492$, $s.e. = 2223$) but positive under the parametric estimator (1222, $s.e. = 1127$), though both are imprecisely estimated due to the small cohort size. While the true effects are unknown, the discrepancy highlights the potential for results to strongly differ based on nuisance function specification. We return to this issue in Section 6.

## 5.2 Dynamic Effects of Hospitalization

In many applications, researchers are primarily interested in summaries of group-time effects. A leading example is the dynamic treatment effect, which underpins many event study designs and provides a way to assess the plausibility of parallel trends.

Formally, dynamic effects are group-weighted averages of the GT-ATTs:

$$\tau_0^{(e)} = \sum_{g \in \mathcal{G}} \mathbb{1}\{g + e \leq T\} \mathrm{P}(G_i = g \mid G_i + e \leq T) \theta_0^{(g,g+e)}, \tag{24}$$

where $\mathcal{G}$ is the set of all treatment initiation periods, and $\theta_0^{(g,t)}$ is the group-time ATT

Table 4: Group-Time Average Treatment Effects on the Treated by Estimator

| Wave first hospitalized | Wave | *Cross-fitting Repetitions* | | | | | Median aggregate | Const. | AIPW |
|---|---|---|---|---|---|---|---|---|---|
| | | Rep. 1 | Rep. 2 | Rep. 3 | Rep. 4 | Rep. 5 | | | |
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 8 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | - | - | - | - | - | - | - | - |
| | 8 | 2467.6 | 2478 | 2442.9 | 2592.9 | 2564.7 | 2478 | 3028.6 | 2200.5 |
| | | (794.9) | (779.6) | (806.4) | (790.5) | (798.4) | (798.8) | (913.5) | (839) |
| | 9 | 626.8 | 466 | 559.2 | 483 | 519.9 | 519.9 | 1247.7 | -14.3 |
| | | (591.9) | (628.3) | (601.6) | (642.8) | (591.9) | (602.8) | (860.7) | (681) |
| | 10 | 723.7 | 803 | 766.6 | 614.8 | 691.8 | 723.7 | 800.1 | 998.9 |
| | | (582.6) | (580.9) | (598.3) | (599.3) | (616.3) | (599.9) | (1007.5) | (570.2) |
| 9 | 7 | 1405.6 | 1358.5 | 1412 | 1333.2 | 1324.4 | 1358.5 | 170 | 1717.8 |
| | | (1022.7) | (997.5) | (992.7) | (969.2) | (965.2) | (994.1) | (1128.4) | (1324.8) |
| | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | - | - | - | - | - | - | - | - |
| | 9 | 3485.6 | 3280.4 | 3218.3 | 3379.2 | 3204.5 | 3280.4 | 3324.4 | 3789.9 |
| | | (1000.4) | (941.9) | (931.1) | (953.2) | (923.3) | (941.9) | (958.8) | (1338.6) |
| | 10 | 983.1 | 936.9 | 1059.8 | 912.5 | 904.6 | 936.9 | 106.8 | 2533.6 |
| | | (431.3) | (427.9) | (432.3) | (436.5) | (425.5) | (433.7) | (650.7) | (421.7) |
| 10 | 7 | -1248 | -1783.6 | -1856.9 | -577.4 | -1491.7 | -1491.7 | 591 | 1221.5 |
| | | (2203.2) | (2473.6) | (2624.7) | (1661.1) | (2223.3) | (2223.3) | (1268.9) | (1126.7) |
| | 8 | 249.5 | 237.9 | 223.2 | 235.4 | 209 | 235.4 | 410.6 | 246.7 |
| | | (1016.2) | (1015.9) | (1019.2) | (1019.2) | (1022.7) | (1019.2) | (1027.1) | (1026.2) |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | - | - | - | - | - | - | - | - |
| | 10 | 2731.1 | 2710.5 | 2690.2 | 2343.6 | 2480.2 | 2690.2 | 3091.5 | 3796.1 |
| | | (1215.2) | (1171.4) | (1260.2) | (1336.2) | (1260.5) | (1260.2) | (995.4) | (931.6) |

*Notes:* Columns (1)-(5) show DML GT-ATT estimators for randomly generated cross-fitting folds. Column (6) presents the corresponding median aggregated DML estimate. Columns (7) and (8) present the Sant'Anna and Zhao (2020) estimator without and with controls, respectively. Point-wise standard errors are in parentheses.

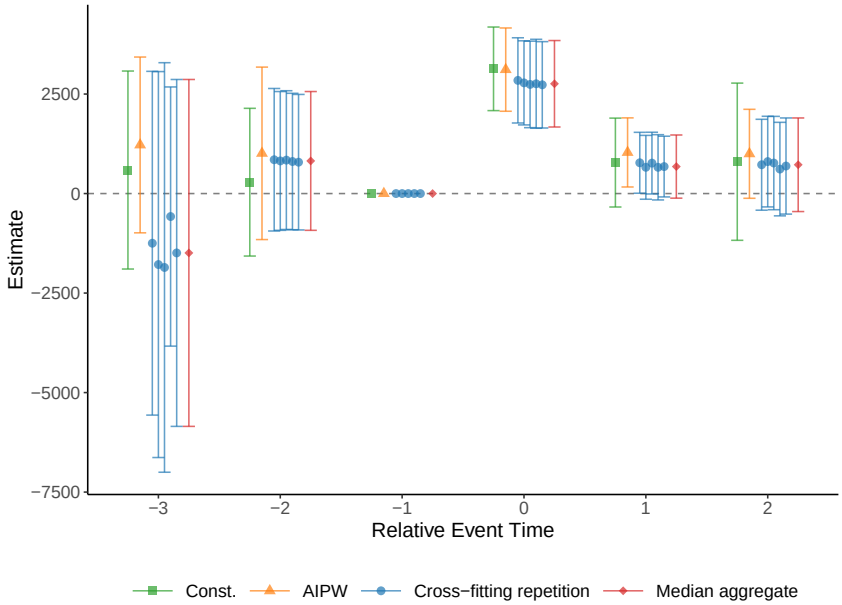
defined earlier. In our context, this corresponds to averaging the effect of hospitalization across all groups observed $e$ periods after their initial admission, weighted by group size. That is, the dynamic effect $\tau_0^{(e)}$ measures the average effect of hospitalization $e$ periods after the initial admission.

Estimating the probabilities $\mathrm{P}\left(G_i = g | G_i + e \leq T\right)$ in (24) is straightforward. For example, a natural estimator is given by the binning estimator $\sum_{i=1} \mathbb{1}\{G_i = g, G_i + e \leq T\} / \sum_{i=1} \mathbb{1}\{G_i + e \leq T\}$. Given an estimator for these probabilities and the DML estimators for the GT-ATTs, inference for $\tau_0^{(e)}$ then follows from standard results about inference on linear combinations.

Figure 3 displays the estimated dynamic effects of hospitalization obtained from the different GT-ATT estimators considered previously. Across methods, we find no statistically significant effects in the pre-treatment period and a large, significant increase in out-of-pocket medical spending immediately following hospitalization. DML estimates and the parametric AIPW alternative are quantitatively similar overall, and the variation in DML across different cross-fitting splits is quite small.

One notable difference is at event time $-3$: the DML estimate is substantially less

Figure 3: Dynamic Average Treatment Effect Estimates




*Notes.* The figure displays dynamic treatment effect estimates using different GT-ATT estimators. "Const." and "AIPW" refer to the Sant'Anna and Zhao (2020) estimator without and with controls, respectively. "Cross-fitting repetition" shows DML GT-ATT estimates across random cross-fitting folds, and "Median aggregate" presents the corresponding median aggregated DML estimate. Bars indicate point-wise 95% confidence intervals.

precise than the parametric alternative. Because only group 10 has outcomes observed three periods before hospitalization, this dynamic effect corresponds exactly to group 10's pre-treatment effect in Wave 7, which we discussed in the preceding section. Since this estimate is based solely on 163 treated and 65 never-treated observations, it is not surprising that it is relatively imprecise. That it is markedly less precise than the corresponding parametric estimate highlights how, with limited data, functional form restrictions can meaningfully shape inference. The discrepancy illustrates how DML can serve as a robustness check on parametric inference when data are limited. At a minimum, it cautions against taking the parametric results at face value without careful economic justification for the functional form assumptions.

We report results for other learners and tuning parameter choices in the companion website for this paper, and find similar results. For example, using random forests with higher regularization (minimum node size of 100) yields a first-period treatment effect estimate of 2903 ($s.e. = 497$), while lower regularization (minimum node size of 1) yields 2666 ($s.e. = 571$). Pre-treatment effects remain statistically insignificant at the 5% level across all learner and tuning parameter choices considered.

The robustness to learner choice is quite different in the example we consider in next section. There we revisit the study on monopsony power of Dube et al. (2020) and illustrate that the selection and tuning of nuisance estimators can be empirically consequential.

We use this application to demonstrate diagnostic analyses that can assist in evaluating and selecting nuisance estimators.

# 6 Monopsony in Online Markets

We revisit the Dube et al. (2020) paper on monopsony power on Amazon MTurk, an online platform for hiring workers to perform tasks. The research falls into the wider literature estimating the labor supply elasticity (see reviews in Sokolova and Sorensen, 2021; Langella and Manning, 2021). This example is useful for at least two reasons. First, it illustrates how DML can incorporate complex, non-tabular data, such as text or images, which are increasingly common in empirical economics (e.g., Ash and Hansen, 2023). Second, it provides an interesting setting for discussing the choice of ML method for nuisance estimation. In particular, it highlights the importance of selecting learners and tuning parameters to support credible inference. We use this example to discuss methods for evaluating such choices.

Dube et al. (2020) consider multiple identification strategies and datasets for studying monopsony power. We focus here on their analysis of a cross-sectional sample compiled by Ipeirotis (2010). The parameter of interest is the partially linear regression coefficient $\theta_0$ introduced in Example 2. The outcome variable, $Y$, is the logarithm of the time required for a posted task to be filled. The treatment variable, $D$, is the logarithm of the payment offered. The covariates, $X$, are a high-dimensional vector of task characteristics.

The authors provide arguments under which $\theta_0$ is the negative labor supply elasticity to the firm and thus a measure of monopsony power. Since tasks vary in complexity, difficulty, and time commitment, it is important to adjust for characteristics that are likely correlated with duration and reward. Some of these task characteristics—such as title, task description, and keywords—come in the form of text. To account for these, the authors include a combination of hand-engineered features and vectorized representations of the task's title, description, and keywords. The hand-engineered features include, among others, the allotted time, common patterns identified using regular expressions, the number of keywords, and the lengths of the title and description. The vectorized text representations used by Dube et al. (2020) consist of Doc2Vec embeddings, topic distributions, and n-grams.

In our reanalysis, we retain the original hand-engineered covariates but replace the text representations used by Dube et al. (2020) with fine-tuned embeddings from the large language model DeBERTa v3. Large language models map text into numerical representations, referred to as embeddings, that capture semantic and contextual information. We use fine-tuning to adapt the pre-trained parameters to our setting, thereby

improving predictive performance.[18] Fine-tuning is performed separately for the outcome and treatment models in each cross-fitting iteration, alleviating concerns about inducing over-fitting bias.

We consider 12 candidate learners that are trained on hand-coded controls and fine-tuned embeddings: OLS as a simple unregularized baseline, cross-validated lasso and ridge as regularized linear methods, and, to allow for nonlinearities, three implementations of random forests, three implementations of gradient boosting (`XGBoost`), and three feed-forward neural network architectures. The tuning parameters are reported in Table 5. We set the number of folds to $K = 3$, repeat cross-fitting $S = 5$ times, and report the median aggregate estimates. Finally, to account for possible dependence in the data, we implement cross-fitting by recruiter, assigning all observations from a given recruiter to the same fold, and compute standard errors clustered at the recruiter level.[19]

Table 5 shows that DML point estimates and standard errors can be highly sensitive to the choice of nuisance estimator, with point estimates ranging from $-3.9$ to 2.1 and standard errors ranging from 0.02 to 3.23. These differences highlight a practical challenge with DML: ML offers a rich set of methods, but different learners may yield qualitatively distinct results. This challenge has been noted in the literature. For example, through calibrated simulations, Ahrens et al. (2025) illustrate stark differences in inferential results for target parameters estimated with DML with different nuisance estimators; see also Wüthrich and Zhu (2023), Giannone, Lenza, and Primiceri (2021), Angrist and Frandsen (2022), and Bach et al. (2024) for related discussions.

A convenient byproduct of cross-fitting is evidence about the out-of-sample performance of nuisance estimators. To provide guidance on the choice of nuisance estimator, we first consider the estimated out-of-sample $R^2$ values calculated using the cross-fitted predicted values and provided in the same table. These values indicate substantial variation across learner specifications. `XGBoost` (columns 7-9) achieves $R^2$ scores of around 85% and 74% for predicting outcome and treatment, respectively, whereas other learners perform markedly worse. The `XGBoost`-based estimates imply a labor supply elasticity between 0.040 and 0.061 ($s.e. \approx 0.022$). By contrast, the neural networks achieve outcome $R^2$ values below 4% and treatment $R^2$ values below 22%, and their corresponding point estimates display considerable instability.

The out-of-sample $R^2$ values provide a useful diagnostic but do not directly indicate whether differences across learners are statistically meaningful. To supplement this, Ta-

[18] We implement fine-tuning using Low-Rank Adaptation (LoRA; Hu et al., 2021). Rather than fully retraining all parameters, LoRA optimizes a low-rank decomposition of the parameters updates, effectively modifying the model within a regularized subspace.

[19] The Online Appendix provides additional implementation details, results for $K = 5$, and results obtained when ignoring dependence in the construction of cross-fitting folds.

Table 5: Estimation results for the Monopsony application: Individual candidate learners

| | *Dependent variable:* log duration | | | | | |
|---|---|---|---|---|---|---|
| *Panel A.* | (1) | (2) | (3) | (4) | (5) | (6) |
| log reward | 0.7548 | −0.0082 | 0.6370 | −0.1938 | −0.1892 | −0.1812 |
| | (2.3904) | (1.3546) | (3.0233) | (0.1344) | (0.1372) | (0.1412) |
| Cross-fitted $R^2$ Outcome | 0.0643 | 0.1471 | 0.0624 | 0.1208 | 0.1169 | 0.1161 |
| Cross-fitted $R^2$ Treatment | 0.2751 | 0.3286 | 0.2834 | 0.4836 | 0.4823 | 0.4817 |
| CVC $p$-value Outcome | 0.0696 | 0.0464 | 0.0672 | 0. | 0. | 0. |
| CVC $p$-value Treatment | 0.1024 | 0.0844 | 0.0528 | 0. | 0. | 0. |
| Weights Outcome | 0. | 0.0007 | 0.0007 | 0. | 0. | 0.0060 |
| Weights Treatment | 0. | 0. | 0. | 0. | 0. | 0. |
| ML | OLS | CV-Lasso | CV-Ridge | RF 1 | RF 2 | RF 3 |
| *Panel B.* | (7) | (8) | (9) | (10) | (11) | (12) |
| log reward | −0.0493 | −0.0397 | −0.0614** | −0.0096 | −3.9293 | 2.0811 |
| | (0.0338) | (0.0221) | (0.0224) | (1.8719) | (3.1104) | (3.2329) |
| Cross-fitted $R^2$ Outcome | 0.8520 | 0.8622 | 0.8636 | 0.0378 | 0.0108 | 0.0140 |
| Cross-fitted $R^2$ Treatment | 0.7366 | 0.7472 | 0.7460 | 0.0260 | 0.1629 | 0.2164 |
| CVC $p$-value Outcome | 0. | 0.2004 | 0.8000 | 0.0892 | 0.0788 | 0.1248 |
| CVC $p$-value Treatment | 0. | 0.4000 | 0.5996 | 0.1000 | 0.0980 | 0.0880 |
| Weights Outcome | 0.2371 | 0.3354 | 0.4159 | 0.0028 | 0.0005 | 0.0010 |
| Weights Treatment | 0.2800 | 0.3626 | 0.3372 | 0.0001 | 0.0138 | 0.0062 |
| ML | XGB 1 | XGB 2 | XGB 3 | NN 1 | NN 2 | NN 3 |

*Notes:* The table reports DML estimates based on 12 different nuisance estimators. The controls are the hand-coded controls of Dube et al. (2020) and DeBERTa embeddings, fine-tuned using LoRA. We use $K = 3$ cross-fitting folds, implemented by randomly assigning recruiters to folds, and report median aggregated estimates obtained using $S = 5$ cross-fitting repetitions. The number of observations is $N = 258,352$. The diagnostics reported are the cross-fitted $R^2$, the CVC $p$-value, and the short-stacking weights; each averaged across cross-fitting repetitions. We consider the following nuisance estimators: OLS; CV-lasso (lasso with tuning parameter selected by cross-validation); CV-ridge (ridge with tuning parameter selected by cross-validation); three types of random forest labeled RF 1 (600 trees), RF 2 (600 trees, minimum terminal node size of 500) and RF 3 (600 trees, minimum terminal node size of 2000); three types of `XGBoost` labeled XGB 1 (800 trees, early stopping after 10 rounds), XGB 2 (800 trees, minimum node size of 500) and XGB 3 (800 trees, minimum node size of 2000). DML estimation uses the `R` package `ddml` (Wiemann et al., 2023). The nuisance estimators were implemented with `glmnet` (Friedman, Hastie, and Tibshirani, 2010), `ranger` (Wright and Ziegler, 2017), `XGBoost` (Chen and Guestrin, 2016), and `keras` (Kalinowski, Allaire, and Chollet, 2025). Standard errors are clustered at the recruiter level.

ble 5 also reports $p$-values from the cross-validation with confidence test (CVC; Lei, 2020). The null hypothesis in the CVC test is that a given learner achieves the lowest predictive risk among the full set of candidates. The alternative is that at least one other learner has a lower risk. Lei (2020) recommends forming a confidence set of the best learners by retaining those for which the $p$-value exceeds a pre-specified threshold, typically 0.1 or 0.2. Using a threshold of 0.2, the CVC tests identify learner specifications 8 and 9 (`XGBoost` 2 and 3) as the only candidate learners for the outcome equation and treatment equation for which the null cannot be rejected.

Rather than selecting the best-performing nuisance estimators based on the $R^2$ score or the CVC test, we can also combine them through stacking or model averaging approaches (Wolpert, 1996; Breiman, 1996; van der Laan, Polley, and Hubbard, 2007). Stacking constructs a "super learner" as a weighted average of candidate learners, where the weights are chosen to minimize out-of-sample prediction error. Within the DML framework, the stacking weights can be re-estimated at each step of the cross-fitting procedure, or

estimated once using the full sample by regressing the outcome on the learners' cross-fitted predicted values (see Ahrens et al. 2025 for a discussion of various options).

In this application, we combine all 12 learners using a method commonly employed in stacking applications, namely constrained least squares: We regress outcome and treatment against cross-fitted predicted values while imposing that the coefficients on the learner predicted values are non-negative and sum to one, and use these estimated coefficients as weights to construct "super learners" for the outcome and treatment. We opt to estimate the stacking regression only once for outcome and treatment on the full sample. The weights are reported in Table 5. The model averaging procedure assigns large weights only to the three `XGBoost` specifications. The procedure produces a DML estimate of $-0.054$ with $s.e. = 0.020$; see Table 6. We obtain similar point estimates when selecting the single best learner for each nuisance function or when assigning equal weights to all learners selected by the CVC test. Our preferred estimates are thus consistent with Dube et al. (2020): using several samples, their DML estimates suggest a labor supply elasticity in the 0.0299—0.198 range.

Table 6: Estimation results for the Monopsony application: Meta learning approaches

| | *Dependent variable:* log duration | | |
|---|---|---|---|
| | Stacking | Single-best | CVC |
| log reward | $-0.0544^{**}$ | $-0.0601^{***}$ | $-0.0589^{*}$ |
| | (0.0198) | (0.0182) | (0.0265) |
| Cross-fitted $R^2$ Outcome | 0.8712 | 0.8643 | 0.8643 |
| Cross-fitted $R^2$ Treatment | 0.7576 | 0.7498 | 0.6620 |

*Notes:* The table reports DML estimates when several learners are combined by constrained least squares (imposing non-negative weights that sum to one), by selecting the single best learner per equation, or by assigning equal weights to all learners for which the CVC $p$-values are larger than 0.2. We use $K = 3$ cross-fitting folds, implemented by randomly assigning recruiters to folds, and report median aggregated estimates obtained using $S = 5$ cross-fitting repetitions. Standard errors are clustered at the recruiter level.

The results in this application highlight that DML estimates can be sensitive to the choice of nuisance estimator. This sensitivity is consistent with conditions in existing DML theory, which indicate that poorly tuned or ill-suited learners can yield misleading results. We emphasize that the choice of which learner to use for a particular nuisance function and data set is rarely, if ever, known *ex ante*. Further, there is no reason to assume that the same learner is best for every nuisance function, although this is implicitly imposed by many common estimation strategies. In practice, we recommend using a diverse set of nuisance function estimators to increase the credibility of DML estimates. Appropriate tools for evaluating these choices include cross-fitted performance metrics such as $R^2$, the CVC test, and model averaging approaches.

# 7 Discussion

Double/Debiased Machine Learning (DML) provides a flexible framework for inference in the presence of high-dimensional nuisance parameters. By combining Neyman orthogonal scores with cross-fitting, it mitigates the impact of estimating nuisance parameters, enabling asymptotically valid inference under relatively weak conditions. These conditions are compatible with a wide range of machine learning methods, making DML particularly useful in applications involving complex non-tabular data such as text or images. More broadly, the ability to leverage flexible estimators makes DML attractive as a complementary robustness check or when researchers wish to avoid parametric assumptions imposed for convenience rather than grounded in economic reasoning.

While DML provides robustness to the estimation of nuisance parameters, it is not a panacea. It does not tell the researcher what interesting target parameters are or how to identify them. Rather, it complements careful reasoning about objects of interest and their identification.

**Implementation Guidance.** When implementing DML, researchers face several design choices. One key decision is how to partition the data. With independent observations, forming folds at random provides a natural default. More generally, researchers should aim to form approximately independent folds by using partitions that respect any underlying dependence structure in the data, as noted in Remark 1.

Researchers also need to select the number of folds $K$. Standard asymptotic results apply for any fixed $K$; see, e.g., Chernozhukov et al. (2018). These results suggest the use of relatively small values for $K$, but do not provide more specific guidance. Velez (2024) provides a higher-order analysis within a restricted class of nuisance estimators and finds that performance improves with more folds, peaking at $K = n$—where $n$ denotes the sample size—but with diminishing returns. Simulation evidence, including that in Section 4, suggests that conventional choices like $K = 5$ or 10 work well in many settings. Given this evidence and that computational complexity is often a concern in applications of DML, we recommend choosing a simple round number informed by available computational resources. When the sample size is very small, larger values of $K$ may be desirable, and if resources permit, sensitivity analysis to $K$ is worthwhile.

To address the algorithmic randomness introduced by sample-splitting, we recommend simply repeating the cross-fitting procedure multiple times and reporting summaries of the resulting estimates, as illustrated in Section 5. This simple procedure reduces unappealing dependence on a particular random split and also serves as a useful diagnostic. Large variation across repetitions may raise concerns about finite-sample behavior or the plausibility of underlying assumptions, while stability across repetitions provides reassur-

ance that results are not sensitive to a particular random split.

Choice of nuisance function estimator is a first-order concern. In Section 6, we showed that results can vary substantially across learners. Because it is rarely clear *ex ante* which learner will perform best in a given application, we recommend considering a diverse set of candidate methods, including parametric benchmarks (e.g., linear or logistic regression), regularized regression, tree-based models, and neural networks. From a practical perspective, greater confidence in DML results is warranted when different learners deliver similar estimates, while substantial divergence across learners should be viewed as a warning sign that calls for caution and further investigation. We illustrated how predictive diagnostics, which are a natural byproduct of cross-fitting, can help identify potentially problematic learners. We encourage researchers to report their candidate learners, tuning procedures, and diagnostic metrics to promote transparency and replicability.

**Challenges and Caveats** While DML is conceptually straightforward, its practical implementation can be challenging, and results can be highly sensitive to implementation choices. Decisions such as how nuisance functions are estimated, how many folds are used, how often cross-fitting is repeated, and how tuning parameters are selected can materially affect empirical results. Developing a deeper understanding of the finite-sample behavior of DML—including tradeoffs in learner complexity, computational complexity, fold count, and cross-fitting repetition—would be valuable not just for refining our theoretical knowledge but for guiding applied practice.

From a theoretical standpoint, part of DML's appeal is that it delivers asymptotically normal inference even when flexible methods are used to estimate nuisance functions. Crucially, provided the nuisance estimator converges sufficiently quickly, the limiting distribution does not depend on the specific choice of nuisance estimator. Of course, it is unrealistic to expect all nuisance estimators to yield sufficiently accurate estimates in all settings. Many modern algorithms do not have theory establishing sufficiently fast rates without strong assumptions. For example, the results in Chi et al. (2022) suggest that random forests, when applied to high-dimensional data with tuning similar to common defaults, may fail to converge quickly enough for DML, and results in Cattaneo, Klusowski, and Yu (2025) establish that deep regression trees may fail to be pointwise consistent. Further work establishing combinations of data-generating processes, learners, and tuning strategies that provide rates of convergence compatible with the sufficient conditions for DML is important for clarifying when, where, and how different learners should be used.

Convergence rate conditions may also fail in settings where nuisance functions are highly complex or non-smooth. Extending DML to better handle such settings is an important direction in current research; see, for example, Robins et al. (2017), Bradic

et al. (2022), and Zheng, Bonvini, and Guo (2025). Relatedly, the literature on sensitivity analysis in DML settings (e.g., Chernozhukov et al., 2024) develops tools that, while motivated by concerns about unobserved confounding, are applicable more broadly. These tools can readily be adapted to assess robustness to other forms of misspecification, including insufficient learner flexibility or excessive nuisance function complexity.

Finally, while accommodating dependent data is conceptually straightforward within the DML framework, its practical implementation raises additional challenges. Forming appropriate partitions, assessing effective sample sizes, and evaluating finite-sample performance become more subtle in dependent settings, further complicating the use of asymptotic approximations as a guide to inference. Given the prevalence of dependent data in economic applications, further development of practical diagnostics and implementation guidance tailored to this setting would be a welcome addition to the literature.

**Functional Restrictions.** A recent proposal for improving performance involves empirical calibration of nuisance function estimates, as developed by van der Laan, Luedtke, and Carone (2025). Their approach adapts ideas from predictive modeling, requiring that estimated nuisance functions satisfy known properties. For example, if the nuisance function is $\eta_0(X) = \mathrm{E}[Y|X]$, we should not be able to improve mean squared prediction error by transforming $\eta_0(X)$. van der Laan, Luedtke, and Carone (2025) propose methods to enforce such calibration and show that, for a broad class of target parameters, DML estimators remain asymptotically normal even if only one nuisance estimator satisfies the usual convergence rate. More generally, using known functional restrictions can reduce the complexity of nuisance estimation, potentially improving performance even in complex settings. More work on incorporating economically motivated constraints seems promising, both as a practical aid and as a direction for methodological research.

**Conclusion.** In sum, while recognizing that important challenges remain, we believe DML is a valuable addition to the empirical researcher's toolkit. It offers a framework for incorporating machine learning methods into empirical analysis, which is increasingly important as economic data grow in richness and complexity. At the same time, part of our goal in this review is to underscore that DML is not a foolproof, mechanical procedure. Empirical results can depend sensitively on implementation choices, and careful diagnostics, robustness checks, and transparent reporting are essential. Viewed in this light, DML should be understood as a structured approach that, when used thoughtfully, can improve empirical practice, either as a mechanism for obtaining primary conclusions from complex data or as a systematic robustness check alongside more traditional methods.

# Appendix

We collect more technical discussions related to orthogonal scores in this appendix. In Appendix A, we illustrate verification of Neyman orthogonality, or the lack thereof, for the scores in Examples 1, 3, and 4. We present a general "partialling out" approach for constructing Neyman orthogonal scores from a given generic score function, which may not itself be Neyman orthogonal, in Appendix B. Finally, we present additional examples of common target parameters and their Neyman orthogonal scores in Appendix C.

# A Verification of Neyman Orthogonality

In this appendix, we verify Neyman orthogonality of the linear regression, linear IV, and AIPW scores presented in Examples 1, 3, and 4. We also verify that the IPW score from Example 4 does not satisfy Neyman orthogonality.

**Example 1 (continued). Neyman Orthogonality of the Linear Regression Score.** In the linear regression score (5), the nuisance parameters $\eta_Y$ and $\eta_D$ are vectors. The second condition in (15) then reduces to

$$\frac{\partial}{\partial \eta_Y}\mathrm{E}[m_{LM}(W;\theta_0,\eta)]\Big|_{\eta=\eta_0} = -\mathrm{E}[X(D - X'\eta_{D,0})] = 0_p$$
$$\frac{\partial}{\partial \eta_D}\mathrm{E}[m_{LM}(W;\theta_0,\eta)]\Big|_{\eta=\eta_0} = 2\theta_0\mathrm{E}[X(D - X'\eta_{D,0})] - \mathrm{E}[X(Y - X'\eta_{Y,0})] = 0_p$$

by applying (4). □

**Example 3 (continued). Neyman Orthogonality of the Linear IV Score.** For the linear IV score (9), the nuisance function is $\eta(Z)$ with true value $\eta_0(Z) = \mathrm{E}[D|Z]$. Considering a small perturbation, $\Delta\eta(Z) = \eta(Z) - \eta_0(Z)$, around $\eta_0(Z)$ yields

$$\frac{\partial}{\partial \lambda}\mathrm{E}[m_{IV}(W;\theta_0,\eta_0(Z) + \lambda\Delta\eta(Z))]\Big|_{\lambda=0} = \mathrm{E}\left[(Y - \theta_0 D)\Delta\eta(Z)\right] = 0,$$

where the last equality follows from $\mathrm{E}[\varepsilon|Z] = 0$.[20] □

**Example 4 (continued). Neyman Orthogonal Scores for the ATE.** We now verify that the IPW score is not Neyman orthogonal, while the AIPW score is.

For the IPW score (11), the nuisance function is $\alpha(D, X)$ with true value $\alpha_0(D, X) =$

[20] Linear IV under conditional mean independence of $Z$ from $\varepsilon$ satisfies a much stronger condition than Neyman orthogonality in that the score equation at $\theta_0$ is globally insensitive to the nuisance function.

$\frac{D}{r_0(X)} - \frac{1-D}{1-r_0(X)}$. Let $\Delta\alpha(D,X) = \alpha(D,X) - \alpha_0(D,X)$, then

$$\frac{\partial}{\partial\lambda}\mathrm{E}\left[m_{IPW}(W_i;\theta_0,\ \alpha_0(D,X) + \lambda\Delta\alpha(D,X))\right]\Big|_{\lambda=0} = \mathrm{E}[\ell_0(D,X)\Delta\alpha(D,X)] \neq 0,$$

where $\ell_0(D,X) = \mathrm{E}[Y|D,X]$. Thus, the IPW score is not Neyman orthogonal.

The AIPW score (12) uses nuisance functions $\eta(D,X) = (\alpha(D,X), \ell(D,X))$ with true values $(\alpha_0(D,X), \ell_0(D,X))$ defined above. Let $\Delta\eta(D,X) = \eta(D,X) - \eta_0(D,X)$, then

$$\begin{aligned}
&\mathrm{E}[m_{AIPW}(W_i;\theta_0, \eta_0(D,X) + \lambda\Delta\eta(D,X))] \\
&\qquad = -\lambda^2\mathrm{E}[\Delta\alpha(D,X)\Delta\ell(D,X)] \\
&\qquad\quad + \mathrm{E}[(\alpha_0(D,X) + \lambda\Delta\alpha(D,X))(Y - \ell_0(D,X))] \\
&\qquad\quad + \lambda\mathrm{E}[(\Delta\ell(1,X) - \Delta\ell(0,X) - \alpha_0(D,X)\Delta\ell(D,X))] \\
&\qquad\quad + \mathrm{E}[\ell_0(1,X) - \ell_0(0,X)] - \theta_0 \\
&\qquad = -\lambda^2\mathrm{E}\left[\Delta\alpha(D,X)\Delta\ell(D,X)\right],
\end{aligned}$$

where the last equality follows from $\mathrm{E}[\ell_0(1,X) - \ell_0(0,X)] - \theta_0 = 0$ by the definition of the ATE, the definition of $\eta_0(D,X)$, and application of the law of expectations. Neyman orthogonality is then immediate:

$$\frac{\partial}{\partial\lambda}\mathrm{E}\left[m(W;\theta_0,\eta_0 + \lambda\Delta\eta)\right]\Big|_{\lambda=0} = \frac{\partial}{\partial\lambda}\lambda^2\Big|_{\lambda=0}\mathrm{E}\left[\Delta\alpha(D,X)\Delta\ell(D,X)\right] = 0.$$

□

# B Constructing Neyman Orthogonal Scores

This appendix illustrates a "partialling out" approach to constructing Neyman orthogonal scores, generalizing the familiar approach of multiple linear regression discussed in Example 1. Throughout, we use $m$ to denote a generic score function and introduce the notation $\psi$ to denote a Neyman orthogonal score. This notational distinction helps clarify the construction of orthogonal scores from baseline moment conditions. We focus on scalar-valued target parameters $\theta_0$ for ease of exposition.

In many settings, we can readily obtain a moment condition $m$ that identifies $\theta_0$ as in (2), i.e., $\mathrm{E}[m(W;\theta_0,\eta_0)] = 0$, but is not Neyman orthogonal. Often, the nuisance parameter in such cases is a vector of conditional expectation functions. That is, $\eta_0 = (\gamma_0^{(h)})_{h=1}^H$ where, for each $h$, $\gamma_0^{(h)}(B^{(h)}) = \mathrm{E}[A^{(h)}|B^{(h)}]$ for some subvectors $A^{(h)}$ and $B^{(h)}$ of $W$. Further, the score function $m$ typically depends on $\eta_0$ only through its value $\eta_0(W)$.

Consider, for example, the ATE defined in (10) with corresponding IPW score

$$m_{IPW}(W;\theta,\eta) = \frac{DY}{\gamma^{(1)}(X)} - \frac{(1-D)Y}{1-\gamma^{(1)}(X)} - \theta,$$

where the nuisance parameter $\eta = \gamma^{(1)}$ takes its true value at $\gamma_0^{(1)}(X) = \mathrm{E}[D|X]$. In terms of the general notation, this corresponds to $A^{(1)} = D$ and $B^{(1)} = X$. Clearly, the IPW score depends on the nuisance parameter only through its value $\gamma_0^{(1)}(X)$.

Newey (1994) shows how to calculate the impact of nuisance estimation for this kind of moment condition and highlights that these calculations facilitate the construction of Neyman orthogonal scores by projecting this impact onto covariates—a generalization of the partialling out approach in multiple linear regression discussed in Example 1. See also Chernozhukov et al. (2018), Chernozhukov et al. (2021), Chernozhukov et al. (2022), and Kennedy (2023a) for further discussion and approaches to construct orthogonal scores.

Heuristically, the first-order impact of bias in the $h^{\text{th}}$ nuisance parameter $\gamma^{(h)}$ is

$$\frac{\partial}{\partial\lambda} m(W;\theta_0,\gamma_0^{(1)},\ldots,\gamma_0^{(h)}+\lambda\Delta\gamma^{(h)},\ldots,\gamma_0^{(H)})\Big|_{\lambda=0} = m_h(W;\theta_0,\eta_0)\Delta\gamma^{(h)}(B^{(h)}) \tag{25}$$

where the equality follows from the chain rule, and we use $\Delta\gamma^{(h)} = \gamma_0^{(h)} - \gamma^{(h)}$ to denote a deviation of the nuisance from its true value and $m_h$ to denote the partial derivative of $m$ with respect to the value of $\gamma^{(h)}$.

To correct for the impact of the bias in the $h^{\text{th}}$ nuisance parameter, define the adjustment factor $\alpha_0^{(h)}(B^{(h)})$ as the projection of $m_h(W;\theta_0,\gamma_0)$ onto covariates $B^{(h)}$:

$$\alpha_0^{(h)}(B^{(h)}) = \mathrm{E}[m_h(W;\theta_0,\eta_0)|B^{(h)}]. \tag{26}$$

By the law of iterated expectations, $\alpha_0^{(h)}$ satisfies the orthogonality condition

$$\mathrm{E}[\alpha_0^{(h)}(B^{(h)})(A^{(h)} - \gamma_0^{(h)}(B^{(h)}))] = 0. \tag{27}$$

Note that the construction of adjustment factors closely parallels that of the best linear predictors in Example 1, which satisfy a weaker—but otherwise analogous—orthogonality condition (4). The key difference is that (26) relaxes the restriction of best *linear* predictors, considering instead the broader class of conditional expectation functions.

A Neyman orthogonal score can then be constructed as

$$\psi(W;\theta,\eta) = m\left(W;\theta,(\gamma^{(h)})_{h=1}^{H}\right) + \sum_{h=1}^{H}\alpha^{(h)}(B^{(h)})(A^{(h)} - \gamma^{(h)}(B^{(h)})), \tag{28}$$

where $\eta = (\gamma^{(h)},\alpha^{(h)})_{h=1}^{H}$ is the combined nuisance parameter with true value $\eta_0 =$

$(\gamma_0^{(h)}, \alpha_0^{(h)})_{h=1}^H$. To verify that the score in (28) indeed satisfies Neyman orthogonality, note first that the orthogonality condition (27) implies that $\psi$ and $m$ identify the same target parameter. Then, by the chain rule

$$\frac{\partial}{\partial\lambda}\Big\{\mathrm{E}\left[\psi(W;\ \theta_0, \eta_0 + \lambda\Delta\eta)\right]\Big\}\Big|_{\lambda=0} = \sum_{h=1}^{H} \mathrm{E}\left[\Big(m_h(W;\theta_0,\gamma_0) - \alpha_0^{(h)}(B^{(h)})\Big)\,\Delta\gamma^{(h)}(B^{(h)})\right] + \sum_{h=1}^{H} \mathrm{E}\left[\Delta\alpha^{(h)}(B^{(h)})\left(A^{(h)} - \gamma_0^{(h)}(B^{(h)})\right)\right] = 0,$$

where the final equality follows from (26) and the law of iterated expectations. Mirroring the familiar approach in multiple linear regression (Example 1) or partially linear regression (Example 2), the score (28) is thus constructed to "partial out" the first-order impact of estimating the nuisance functions: for each $h$, small errors in either $\gamma_0^{(h)}(B^{(h)})$ or $\alpha_0^{(h)}(B^{(h)})$ can be offset by the other, making estimation more robust.

Returning to the example of the average treatment effect and its IPW score, we can thus compute the impact of propensity score estimation as

$$\frac{\partial}{\partial\lambda} m(W;\theta_0,\gamma_0^{(1)} + \lambda\Delta\gamma^{(1)})\Big|_{\lambda=0} = m_1(W;\theta_0,\gamma_0^{(1)})\Delta\gamma^{(1)}(X),$$

where $m_1(W;\theta_0,\gamma_0^{(1)}) = -\frac{DY}{\gamma_0^{(1)}(X)^2} - \frac{(1-D)Y}{(1-\gamma_0^{(1)}(X))^2}$. We then construct the adjustment factor as

$$\alpha_0^{(1)}(X) = \mathrm{E}[m_1(W;\theta_0,\gamma_0^{(1)})|X] = -\frac{\mathrm{E}[Y|D=1,X]}{\gamma_0^{(1)}(X)} - \frac{\mathrm{E}[Y|D=0,X]}{1-\gamma_0^{(1)}(X)},$$

where the final equality follows from the law of total probability. Combining provides a Neyman orthogonal score for the ATE:

$$\psi(W;\theta,\eta) = \frac{DY}{\gamma^{(1)}(X)} - \frac{(1-D)Y}{1-\gamma^{(1)}(X)} - \theta + \alpha^{(1)}(X)(D - \gamma^{(1)}(X)).$$

After replacing $\alpha^{(1)}$ with its representation in terms of conditional expectations and some algebra, this expression reduces to the AIPW score for the ATE in Example 4.

**Remark 4** (Adjustment Factors $\alpha_0^{(h)}$ are Riesz Representers)
The orthogonal score construction described above is closely connected to the concept of the *Riesz representer*, a fundamental object in functional analysis. In particular, the adjustment factors (26) are Riesz representers for the expected first-order impact of bias in the $h^{\text{th}}$ nuisance function. To see this, note that taking expectations of the

first-order bias (25) defines a functional $\phi^{(h)} : \Gamma^{(h)} \to \mathbb{R}$ for each $h$,

$$\phi^{(h)}(f) \equiv \mathrm{E}\left[m_h(W;\theta_0,\eta_0)f(B^{(h)})\right], \tag{29}$$

where $\Gamma^{(h)}$ denotes the appropriate functional space of $\gamma_0^{(h)}$ and $f \in \Gamma^{(h)}$ denotes any potential nuisance bias function $\Delta\gamma^{(h)}$. Under appropriate conditions, the Riesz representation theorem implies existence of a unique $\alpha_0^{(h)} \in \Gamma^{(h)}$ such that for any $f \in \Gamma^{(h)}$, $\phi^{(h)}(f) = \mathrm{E}[\alpha_0^{(h)}(W)f(B^{(h)})]$. The law of iterated expectations then implies that these so-called Riesz representers are given by the adjustment factors (26).

The insight that the adjustment factors are Riesz representers facilitates construction of Neyman orthogonal scores, even when an explicit expression for $\alpha_0$ is unavailable. This is key to recent DML approaches that rely on implicitly defined Riesz representers (e.g., Chernozhukov et al., 2021; Chernozhukov, Newey, and Singh, 2022a; Chernozhukov, Newey, and Singh, 2022b; Hirshberg and Wager, 2021).

**Remark 5** (Calibrated Riesz Representers)
Viewing the adjustment factors $\alpha_0^{(h)}$ as Riesz representers as highlighted in Remark 4 allows researchers to leverage properties implied by the Riesz representation theorem that can further improve finite sample behavior of DML estimators. As an example, note that taking $f = \alpha_0^{(h)}$ in (29) where $\alpha_0^{(h)} \in \Gamma^{(h)}$, we have $\mathrm{E}\left[m_h(W;\theta_0,\gamma_0)\alpha_0^{(h)}(B^{(h)})\right] = \mathrm{E}[\alpha_0^{(h)}(B^{(h)})^2]$. This moment condition holds in the population, but its sample analog based on the DML estimator in Algorithm 1 generally does not. To ensure that the sample analog holds as well, we can calibrate an initial estimate of the adjustment factor $\widehat{\alpha}^{(h)}$ by introducing a scalar $\widehat{\iota}_h$ and defining $\widehat{\alpha}_{\mathrm{cal}}^{(h)} = \widehat{\iota}_h\widehat{\alpha}^{(h)}$. Assuming $\widehat{\iota}_h \neq 0$, $\widehat{\iota}_h$ is set to

$$\mathbb{E}_n\left[\widehat{m}_h(W)(\widehat{\iota}_k\widehat{\alpha}^{(h)}(B^{(h)}))\right] = \mathbb{E}_n\left[(\widehat{\iota}_h\widehat{\alpha}^{(h)}(B^{(h)}))^2\right] \Leftrightarrow \widehat{\iota}_h = \frac{\mathbb{E}_n[\widehat{m}_h(B^{(h)})\widehat{\alpha}^{(h)}(B^{(h)})]}{\mathbb{E}_n[\widehat{\alpha}^{(h)}(B^{(h)})^2]}.$$

See Laan, Luedtke, and Carone (2025) for further discussion of *calibrated* DML.

# C Neyman Orthogonal Scores for Additional Common Target Parameters

We now provide additional examples of common target parameters and their Neyman orthogonal scores. For ease of exposition, we categorize these parameters as treatment effect

parameters (Section C.1), regression parameters (Section C.2), and fixed effect regression parameters (Section C.3). We use the same notation and structure as in Appendix B.

## C.1 Treatment Effect Parameters

Consider $W = (Y, D, Z, X)$ where $Y$ is a scalar outcome, $D$ is a discrete treatment, $Z$ is a binary instrument, and $X$ is a vector of controls. Under standard assumptions (e.g., Imbens and Rubin, 2015), the following objects are well-defined causal quantities.

### C.1.1 Weighted Average Potential Outcome

The weighted average potential outcome corresponding to treatment level $d$ is

$$\theta_0 = \mathrm{E}[\omega(X)\mathrm{E}[Y|D = d, X]],$$

for some known weighting function $\omega$. The corresponding IPW score is

$$m(W; \theta, \gamma^{(1)}) = \frac{\mathbb{1}\{D = d\}Y}{\gamma^{(1)}(X)}\omega(X) - \theta,$$

where the nuisance parameter $\gamma^{(1)}$ has true value $\gamma_0^{(1)}(X) = \mathrm{E}[\mathbb{1}\{D = d\}|X]$. The corresponding correction term is given by

$$\alpha_0^{(1)}(X) = -\frac{\mathrm{E}[Y|D = d, X]\omega(X)}{\mathrm{E}[\mathbb{1}\{D = d\}|X]}.$$

### C.1.2 Average Treatment Effect on the Treated

The average treatment effect on the treated for a binary treatment $D$ is

$$\theta_0 = \mathrm{E}\left[\mathrm{E}[Y|D = 1, X] - \mathrm{E}[Y|D = 0, X]|D = 1\right],$$

with corresponding IPW score

$$m(W; \theta, \gamma^{(1)}, \gamma^{(2)}) = \frac{DY}{\gamma^{(2)}} - \frac{\gamma^{(1)}(X)(1 - D)Y}{\gamma^{(2)}(1 - \gamma^{(1)}(X))} - \frac{D}{\gamma^{(2)}}\theta,$$

where the nuisance parameters $\gamma^{(1)}$ and $\gamma^{(2)}$ take true values at $\gamma_0^{(1)}(X) = \mathrm{E}[D|X]$ and $\gamma_0^{(2)} = \mathrm{E}[D]$. The corresponding correction terms are given by

$$\alpha_0^{(1)}(X) = -\frac{1}{\mathrm{E}[D]}\left(\frac{\mathrm{E}[D|X]\mathrm{E}[Y|D = 0, X]}{1 - \mathrm{E}[D|X]} + \mathrm{E}[Y|D = 0, X]\right), \quad \alpha_0^{(2)} = 0.$$

### C.1.3 Local Average Treatment Effect

The local average treatment effect (LATE) for a binary instrument $Z$ is defined as

$$\theta_0 = \frac{\mathrm{E}\left[\mathrm{E}[Y|Z=1,X] - \mathrm{E}[Y|Z=0,X]\right]}{\mathrm{E}\left[\mathrm{E}[D|Z=1,X] - \mathrm{E}[D|Z=0,X]\right]},$$

with corresponding IPW score

$$m(W;\theta,\gamma^{(1)}) = \frac{ZY}{\gamma^{(1)}(X)} - \frac{(1-Z)Y}{1-\gamma^{(1)}(X)} - \theta\left(\frac{ZD}{\gamma^{(1)}(X)} - \frac{(1-Z)D}{1-\gamma^{(1)}(X)}\right),$$

where the nuisance parameter $\gamma^{(1)}$ takes true value at $\gamma_0^{(1)}(X) = \mathrm{E}[Z|X]$. The corresponding correction term is given by

$$\alpha_0^{(1)}(X) = -\frac{\mathrm{E}[Y|Z=1,X]}{\mathrm{E}[Z|X]} - \frac{\mathrm{E}[Y|Z=0,X]}{1-\mathrm{E}[Z|X]} + \theta_0\left(\frac{\mathrm{E}[D|Z=1,X]}{\mathrm{E}[Z|X]} + \frac{\mathrm{E}[D|Z=0,X]}{1-\mathrm{E}[Z|X]}\right).$$

## C.2 Regression Parameters

Consider $W = (Y, D, Z, X)$ where $Y$ is a scalar-valued outcome, $D$ is a vector of variables of interest, $Z$ is a vector of instruments, and $X$ is a vector of controls.

### C.2.1 Partially Linear Regression and Partially Linear IV

Consider the instrumental variable (IV) regression

$$Y = D^\top\theta_0 + g_0(X) + \varepsilon,$$

where the target parameter $\theta_0$ and confounding function $g_0(\cdot)$ are defined through the orthogonality restrictions $\mathrm{E}[Z\varepsilon] = 0$ and $\mathrm{E}[\varepsilon|X] = 0$, and the IV relevance condition that $\mathrm{E}[\mathrm{Cov}(Z, D|X)]$ has full column rank. This setting corresponds to a scenario where a researcher has a known set of instruments, $Z$, that are taken to satisfy the exclusion restriction only after conditioning on controls $X$ and does not wish to impose the functional form in which confounds enter the model. Further note that we recover partially linear regression by setting $Z = D$. Solving for $g_0$ and substituting, we obtain the vector-valued score

$$m(W;\theta,\gamma^{(1)},\gamma^{(2)}) = Z(Y - \gamma^{(1)}(X) - \theta^\top(D - \gamma^{(2)}(X))),$$

where the nuisance parameters take true value at $\gamma_0^{(1)}(X) = \mathrm{E}[Y|X]$ and $\gamma_0^{(2)}(X) = \mathrm{E}[D|X]$. The corresponding correction terms are given by

$$\alpha_0^{(1)}(X) = -\mathrm{E}[Z|X], \quad \alpha_0^{(2)}(X) = \mathrm{E}[Z|X]\theta^\top.$$

### C.2.2 Flexible Partially Linear Instrumental Variables

Consider the IV regression

$$Y = D^\top\theta_0 + g_0(X) + \varepsilon,$$

where the target parameter $\theta_0$ and the confounding function $g_0(\cdot)$ are defined through the orthogonality restrictions $\mathrm{E}[\varepsilon|Z,X] = 0$, and the IV relevance condition that $\mathrm{E}[\mathrm{Var}(\mathrm{E}[D|Z,X]|X)]$ is positive definite. This setting differs from that considered in Section C.2.1 in that the orthogonality condition $\mathrm{E}[\varepsilon|Z,X] = 0$ is stronger. It implies that any function of $(Z,X)$ can be used as a valid instrument. We consider the optimal instrument under homoskedasticity ($\mathrm{E}[D|Z,X]$).

Using the fact that $\mathrm{E}[\varepsilon|Z,X] = 0$ implies $\mathrm{E}[\varepsilon|X] = 0$ by the law of iterated expectations, we can solve for $g_0$. Further, note that $\mathrm{E}[\varepsilon|Z,X] = 0$ implies $\mathrm{E}[\gamma_0^{(3)}(Z,X)\varepsilon] = 0$ for $\gamma_0^{(3)}(Z,X) = \mathrm{E}[D|Z,X]$. This motivates the vector-valued score

$$m(W;\theta,\gamma^{(1)},\gamma^{(2)},\gamma^{(3)}) = \gamma^{(3)}(Z,X)(Y - \gamma^{(1)}(X) - \theta^\top(D - \gamma^{(2)}(X))),$$

where the nuisance parameters $\gamma^{(1)}$, $\gamma^{(2)}$, and $\gamma^{(3)}$ take true values, respectively, at $\gamma_0^{(1)}(X) = \mathrm{E}[Y|X]$, $\gamma_0^{(2)}(X) = \mathrm{E}[D|X]$, and $\gamma_0^{(3)}(Z,X) = \mathrm{E}[D|Z,X]$. The corresponding correction terms are given by

$$\alpha_0^{(1)}(X) = -\mathrm{E}[D|X], \quad \alpha_0^{(2)}(X) = \mathrm{E}[D|X]\theta^\top, \quad \alpha_0^{(3)}(Z,X) = 0.$$

## C.3 Fixed Effect Regression Parameters

Consider $W_i = (Y_{i,t}, D_{i,t}, Z_{i,t}, X_{i,t})_{t=0}^T$ where $t$ denotes a secondary dimension (e.g., time), $Y_{i,t}$ is a scalar-valued outcome, $D_{i,t}$ is a vector of variables of interest, $Z_{i,t}$ is a vector of instruments, and $X_{i,t}$ is a vector of controls. We explicitly index by $i$ and $t$ to introduce cross-sectional heterogeneity via individual fixed effects. It is convenient to define the first difference operator $\Delta A_{i,t} = A_{i,t} - A_{i,t-1}$ for random variables $A_{i,t}$ and $A_{i,t-1}$.

Consider the IV regression with fixed effects $\iota_i$

$$Y_{i,t} = D_{i,t}^\top\theta_0 + g_0^{(t)}(X_{i,t}) + \iota_i + \varepsilon_{i,t}, \quad \forall t = 0,1,\ldots,T,$$

where the target parameter $\theta_0$ and the differenced confounding functions $\{\Delta g_0^{(t)}\}_{t=0}^T$ are defined through the orthogonality restrictions $\mathrm{E}[\sum_{t=1}^T \Delta Z_{i,t}\Delta\varepsilon_{i,t}] = 0$, $\mathrm{E}[\Delta\varepsilon_{i,t}|X_{i,t}, X_{i,t-1}] = 0, \forall t \in \{1,\dots,T\}$, and the IV relevance condition that $\mathrm{E}[\mathrm{Cov}(\Delta Z_{i,t}, \Delta D_{i,t}|X_{i,t}, X_{i,t-1})]$ has full column rank for at least some $t \in \{1,\dots,T\}$. Note that we recover fixed effects partially linear regression by setting $(Z_{i,t})_{t=1}^T = (D_{i,t})_{t=1}^T$. Solving for $\Delta g_0^{(t)}$ and substituting, we obtain the vector-valued score

$$m(W_i;\theta,\{\gamma^{(1t)},\gamma^{(2t)}\}_{t=1}^T) = \sum_{t=1}^T \Delta Z_{i,t}(\Delta Y_{i,t} - \gamma^{(1t)}(X_{i,t},X_{i,t-1}) - \theta^\top(\Delta D_{i,t} - \gamma^{(2t)}(X_{i,t},X_{i,t-1}))),$$

where the nuisance parameters $\{\gamma^{(1t)},\gamma^{(2t)}\}_{t=1}^T$ take true value at $\gamma_0^{(1t)}(X_{i,t},X_{i,t-1}) = \mathrm{E}[\Delta Y_{i,t}|X_{i,t},X_{i,t-1}]$ and $\gamma_0^{(2t)}(X_i) = \mathrm{E}[\Delta D_{i,t}|X_{i,t},X_{i,t-1}]$, for all $t \in \{1,\dots,T\}$. The corresponding correction terms are given by

$$\alpha_0^{(1t)}(X_{i,t},X_{i,t-1}) = -\mathrm{E}[\Delta Z_{i,t}|X_{i,t},X_{i,t-1}], \quad \alpha_0^{(2t)}(X_{i,t},X_{i,t-1}) = \mathrm{E}[\Delta Z_{i,t}|X_{i,t},X_{i,t-1}]\theta^\top.$$

# References


Abadie, Alberto, Anish Agarwal, Raaz Dwivedi, and Abhin Shah (2024). "Doubly Robust Inference in Causal Latent Factor Models". *arXiv:2402.11652*.

Ahrens, Achim, Christian B. Hansen, Mark E. Schaffer, and Thomas Wiemann (2024). "ddml: Double/debiased machine learning in Stata". *The Stata Journal* 24.1, pp. 3–45.

— (2025). "Model averaging and double machine learning". *Journal of Applied Econometrics* 40.3, pp. 249–269.

Andrews, Donald W. K. (1994). "Asymptotics for semiparametric econometric models via stochastic equicontinuity". *Econometrica* 62.1, pp. 43–72.

Angrist, J. D. and A. B. Krueger (1995a). "Split-Sample Instrumental Variables Estimates of the Return to Schooling". *Journal of Business & Economic Statistics* 13.2, pp. 225–235.

Angrist, Joshua D. and Brigham Frandsen (2022). "Machine labor". *Journal of Labor Economics* 40.S1, S97–S140.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger (1999). "Jackknife Instrumental Variables Estimation". *Journal of Applied Econometrics* 14.1, pp. 57–67.

Angrist, Joshua D. and Alan B. Krueger (1995b). "Split-Sample Instrumental Variables Estimates of the Return to Schooling". *Journal of Business and Economic Statistics* 13, pp. 225–235.

Angrist, Joshua D and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ash, Elliott and Stephen Hansen (2023). "Text Algorithms in Economics". *Annual Review of Economics* 15.1, annurev–economics–082222–074352.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019). "Ensemble Methods for Causal Effects in Panel Data Settings". *AEA Papers and Proceedings* 109, pp. 65–70.

Athey, Susan and Guido W. Imbens (2019). "Machine learning methods that economists should know about". *Annual Review of Economics* 11.1, pp. 685–725.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". *Annals of Statistics* 47.2, pp. 1148–1178.

Athey, Susan and Stefan Wager (2021). "Policy Learning With Observational Data". *Econometrica* 89.1, pp. 133–161.

Bach, Philipp, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler (2021). *DoubleML – An Object-Oriented Implementation of Double Machine Learning in R*.

— (2022). "DoubleML – An Object-Oriented Implementation of Double Machine Learning in Python". *Journal of Machine Learning Research* 23.53, pp. 1–6.

Bach, Philipp, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler (2024). "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study". In: *Proceedings of the Third Conference on Causal Learning and Reasoning*. PMLR, 236:1065–1117.

Baker, Andrew C, David F Larcker, and Charles CY Wang (2022). "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144.2, pp. 370–395.

Ballinari, Daniele and Alexander Wehrli (2025). "Semiparametric inference for impulse response functions using double/debiased machine learning". *arXiv:2411.10009*.

Bekker, Paul A. (1994). "Alternative Approximations to the Distributions of Instrumental Variables Estimators". *Econometrica* 63, pp. 657–681.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". *Econometrica* 80.6, pp. 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen (2017). "Program evaluation and causal inference with high-dimensional data". *Econometrica* 85.1, pp. 233–298.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2010). "Lasso methods for gaussian instrumental variables models". *arXiv:1012.1297*.

— (2014). "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls". *Review of Economic Studies* 81.2, pp. 608–650.

Bia, Michela, Martin Huber, and Lukáš Lafférs (2024). "Double Machine Learning for Sample Selection Models". *Journal of Business & Economic Statistics* 42.3, pp. 958–969.

Bickel, Peter J. (1982). "On Adaptive Estimation". *Annals of Statistics* 10, pp. 647–671.

Bickel, Peter J. and Ya'acov Ritov (1988). "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates". *Sankhya A* 50.3, pp. 381–393.

Bickel, Peter J, Ya'acov Ritov, and Alexandre B Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector". *Annals of Statistics* 37.4, pp. 1705–1732.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2022). "When is TSLS Actually LATE?" *BFI Working Paper* 2022-16.

Bonvini, Matteo, Alec McClean, Zach Branson, and Edward H. Kennedy (2021). "Incremental Causal Effects: An Introduction and Review". *arXiv:2110.10532*.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024). "Revisiting event study designs: Robust and efficient estimation". *Review of Economic Studies* 91.6, 3253–3285.

Bradic, Jelena, Victor Chernozhukov, Whitney K. Newey, and Yinchu Zhu (2022). "Minimax Semiparametric Learning With Approximate Sparsity". *arXiv:1912.12213*.

Breiman, Leo (1996). "Stacked regressions". *Machine Learning* 24.1, pp. 49–64.

Callaway, Brantly and Pedro H. C. Sant'Anna (2021). "Difference-in-differences with multiple time periods". *Journal of Econometrics* 225 (2), pp. 200–230.

Cattaneo, Matias D., Jason M. Klusowski, and Ruiqi Rae Yu (2025). "The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation". *arXiv:2509.11381*.

de Chaisemartin, Clément and Xavier d'Haultfoeuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9, pp. 2964–2996.

— (2023). "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey". *Econometrics Journal* 26.3, pp. C1–C30.

Chang, Neng-Chieh (2020). "Double/debiased machine learning for difference-in-differences models". *Econometrics Journal* 23.2, pp. 177–191.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen (2012). "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments". *Econometric Theory* 28.1, pp. 42–86.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, USA: ACM, pp. 785–794.

Chen, Xiaohong (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". In: *Handbook of Econometrics*. Ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Elsevier. Chap. 76, pp. 5549–5632.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). "Double/debiased machine learning for treatment and structural parameters". *Econometrics Journal* 21.1, pp. C1–C68.

Chernozhukov, Victor, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis (2024). "Long Story Short: Omitted Variable Bias in Causal Machine Learning". *arXiv:2112.13398*.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins (2022). "Locally robust semiparametric estimation". *Econometrica* 90.4, pp. 1501–1535.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (2015). "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach". *Annual Review of Economics* 7.1.

Chernozhukov, Victor, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis (2021). "Automatic debiased machine learning via neural nets for generalized linear regression". *arXiv:2104.14737*.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (2022a). "Automatic debiased machine learning of causal and structural effects". *Econometrica* 90.3, pp. 967–1027.

— (2022b). "Debiased machine learning of global and local parameters using regularized Riesz representers". *Econometrics Journal* 25.3, pp. 576–601.

Chi, Chien-Ming, Patrick Vossler, Yingying Fan, and Jinchi Lv (2022). "Asymptotic properties of high-dimensional random forests". *Annals of Statistics* 50.6, pp. 3415–3438.

Chiang, Harold D, Kengo Kato, Yukun Ma, and Yuya Sasaki (2022). "Multiway cluster robust double/debiased machine learning". *Journal of Business & Economic Statistics* 40.3, pp. 1046–1056.

Chiang, Harold D, Yukun Ma, Joel B Rodrigue, and Yuya Sasaki (2026). "Double/Debiased Machine Learning for Dyadic Data". *Econometric Theory*, forthcoming.

Chyn, Eric, Brigham Frandsen, and Emily C Leslie (2024). "Examiner and Judge Designs in Economics: A Practitioner's Guide". *Journal of Economic Literature* forthcoming.

Clarke, Paul S. and Annalivia Polselli (2024). *Double Machine Learning for Static Panel Models with Fixed Effects*.

Colangelo, Kyle and Ying-Ying Lee (2023). "Double debiased machine learning nonparametric inference with continuous treatments". *arXiv:2004.03036*.

Deaner, Ben (2023). "Many Proxy Controls". *arXiv:2110.03973*.

Dell, Melissa (2024). "Deep Learning for Economists". *Journal of Economic Literature* forthcoming.

Díaz, Iván (2020). "Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning". *Biostatistics* 21.2, pp. 353–358.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J Notowidigdo (2018). "The economic consequences of hospital admissions". *American Economic Review* 108.2, pp. 308–352.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri (2020). "Monopsony in Online Labor Markets". *American Economic Review: Insights* 2.1, pp. 33–46.

Dustmann, Christian, Francesco Fasani, and Biagio Speciale (2017). "Illegal Migration and Consumption Behavior of Immigrant Households". *Journal of the European Economic Association* 15.3, pp. 654–691.

Escanciano, Juan Carlos and Telmo Pérez-Izquierdo (2023). "Automatic Locally Robust Estimation with Generated Regressors". *arXiv:2301.10643*.

Fan, Jianqing, Shaojun Guo, and Ning Hao (2012). "Variance estimation using refitted cross-validation in ultrahigh dimensional regression". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 74.1, pp. 37–65.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021a). "Deep Learning for Individual Heterogeneity: An Automatic Inference Framework". *arXiv:2010.14694*.

— (2021b). "Deep Neural Networks for Estimation and Inference". *Econometrica* 89.1, pp. 181–213.

Foster, Dylan J. and Vasilis Syrgkanis (2023). "Orthogonal statistical learning". *Annals of Statistics* 51.3, pp. 879–908.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software* 33.1, pp. 1–22.

van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". *Annals of Statistics* 42.3, pp. 1166–1202.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". *Journal of Economic Literature* 57.3, pp. 535–574.

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2021). "Economic predictions with big data: The illusion of sparsity". *Econometrica* 89.5, pp. 2409–2437.

Gilchrist, Duncan Sheppard and Emily Glassberg Sands (2016). "Something to talk about: Social spillovers in movie consumption". *Journal of Political Economy* 124.5, pp. 1339–1382.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024). "Contamination bias in linear regressions". *American Economic Review* 114.12, pp. 4015–4051.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225.2, pp. 254–277.

Haddad, Michel F. C., Martin Huber, and Lucas Z. Zhang (2024). "Difference-in-Differences with Time-varying Continuous Treatments using Double/Debiased Machine Learning". *arXiv:2410.21105*.

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

Hansen, Christian and Damian Kozbur (2014). "Instrumental variables estimation with many weak instruments using regularized JIVE". *Journal of Econometrics* 182.2, pp. 290–308.

Hasminskii, Rafail Z. and Ildar A. Ibragimov (1978). "On the nonparametric estimation of functionals". In: *Proceedings 2nd Prague Symposium on Asymptotic Statistics*. North-Holland, pp. 41–51.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. 2nd ed. New York: Springer-Verlag.

Heckman, James J and Edward J Vytlacil (2007). "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation". In: *Handbook of Econometrics*. Ed. by James J Heckman and E Leamer. Vol. 6. Amsterdam: Elsevier. Chap. 70, pp. 4779–4874.

Hidalgo, F. Daniel, Suresh Naidu, Simeon Nichter, and Neal Richardson (2010). "Economic Determinants of Land Invasions". *Review of Economics and Statistics* 92.3, pp. 505–523.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt (2022). "Demystifying statistical learning based on efficient influence functions". *The American Statistician* 76.3, pp. 292–304.

Hirshberg, David A. and Stefan Wager (2021). "Augmented minimax linear estimation". *Annals of Statistics* 49.6, pp. 3206 –3227.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.

Hubbard, Alan E., Sara Kherad-Pajouh, and Mark J. van der Laan (2016). "Statistical Inference for Data Adaptive Target Parameters". *International Journal of Biostatistics* 12.1, pp. 3–19.

Ichimura, Hidehiko and Petra E. Todd (2007). "Implementing Nonparametric and Semiparametric Estimators". In: *Handbook of Econometrics*. Ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Elsevier. Chap. 74, pp. 5369–5468.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ipeirotis, Panagiotis G (2010). "Analyzing the amazon mechanical turk marketplace". *XRDS: Crossroads, The ACM magazine for students* 17.2, pp. 16–21.

James, G, D Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Cham.

Javanmard, Adel and Andrea Montanari (2014). "Confidence intervals and hypothesis testing for high-dimensional regression". *Journal of Machine Learning Research* 15.1, pp. 2869–2909.

Jung, Yonghan, Jin Tian, and Elias Bareinboim (2021). "Double Machine Learning Density Estimation for Local Treatment Effects with Instruments". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 21821–21833.

Kalinowski, Tomasz, JJ Allaire, and François Chollet (2025). *keras3: R Interface to 'Keras'*.

Kallus, Nathan, Xiaojie Mao, and Masatoshi Uehara (2024). "Localized Debiased Machine Learning: Efficient Inference on Quantile Treatment Effects and Beyond". *Journal of Machine Learning Research* 25.16, pp. 1–59.

Kennedy, Edward H. (2023a). "Semiparametric doubly robust targeted double machine learning: a review". *arXiv:2203.06469*.

— (2023b). "Towards optimal doubly robust estimation of heterogeneous causal effects". *Electronic Journal of Statistics* 17.2, pp. 3008–3049.

Klosin, Sylvia (2021). "Automatic Double Machine Learning for Continuous Treatment Effects". *arXiv:2104.10334*.

Klosin, Sylvia and Max Vilgalys (2023). "Estimating Continuous Treatment Effects in Panel Data using Machine Learning with a Climate Application". *arXiv:2207.08789*.

van der Laan, Lars, Alex Luedtke, and Marco Carone (2025). "Doubly Robust Inference via Calibration". *arXiv:2411.02771*.

Laan, Lars van der, Alex Luedtke, and Marco Carone (2025). "Doubly robust inference via calibration". *arXiv:2411.02771*.

van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard (2007). "Super Learner". *Statistical Applications in Genetics and Molecular Biology* 6.

van der Laan, Mark J. and Sherri Rose (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

van der Laan, Mark J. and Donald Rubin (2006). "Targeted maximum likelihood learning", Working Paper 213, UC Berkeley Division of Biostatistics Working Paper Series.

Langella, Monica and Alan Manning (2021). "Marshall Lecture 2020: The Measure of Monopsony". *Journal of the European Economic Association* 19.6, pp. 2929–2957.

Lei, Jing (2020). "Cross-Validation With Confidence". *Journal of the American Statistical Association* 115.532, pp. 1978–1997.

Lewis, Greg and Vasilis Syrgkanis (2021). "Double/Debiased Machine Learning for Dynamic Treatment Effects". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 22695–22707.

Li, Qi and Jeffrey Scott Racine (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press: Princeton, NJ.

Ma, Yukun (2023). "Identification-robust inference for the LATE with high-dimensional covariates". *arXiv:2302.09756*.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". *Journal of Economic Perspectives* 31.2, pp. 87–106.

Newey, Whitney K. (1994). "The asymptotic variance of semiparametric estimators". *Econometrica* 62.6, pp. 1349–1382.

Newey, Whitney K. and Daniel McFadden (1994). "Large Sample Estimation and Hypothesis Testing". In: *Handbook of Econometrics. Volume 4*. Ed. by R. F. Engle and D. L. McFadden. Elsevier: North-Holland.

Neyman, Jerzy (1959). "Optimal asymptotic tests of composite hypotheses". *Probability and statsitics*, pp. 213–234.

— (1979). "$C(\alpha)$ tests and their use". *Sankhya* 41, pp. 1–21.

Nie, Xinkun and Stefan Wager (2021). "Quasi-oracle estimation of heterogeneous treatment effects". *Biometrika* 108.2, pp. 299–319.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*. New York: Springer.

Poterba, James M, Steven F Venti, and David A Wise (1995). "Do 401(k) contributions crowd out other personal saving?" *Journal of Public Economics* 58.1, pp. 1–32.

Robins, James M, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart (2017). "Minimax estimation of a functional on a structured high-dimensional model". *Annals of Statistics* 45.5, pp. 1951–1987.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American Statistical Association* 89.427, pp. 846–866.

Robins, James M, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, Aad van der Vaart, and HEI Health Review Committee (2013). "New statistical approaches to semiparametric regression with application to air pollution research". *Research report (Health Effects Institute)* 175, pp. 3–129.

Robins, James, Lingling Li, Eric Tchetgen, and Aad van der Vaart (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman*. Vol. 2. Institute of Mathematical Statistics, pp. 335–422.

Robinson, Peter M. (1988). "Root-$N$-consistent semiparametric regression". *Econometrica* 56.4, pp. 931–954.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe (2023). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature". *Journal of Econometrics* 235.2, pp. 2218–2244.

Sant'Anna, Pedro H. C. and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". *Journal of Econometrics* 219.1, pp. 101–122.

Sasaki, Yuya and Takuya Ura (2023). "Estimation and inference for policy relevant treatment effects". *Journal of Econometrics* 234.2, pp. 394–450.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins (1999). "Rejoinder to "Adjusting for non-ignorable drop-out using semiparametric non-response models"". *Journal of the American Statistical Association* 94, pp. 1135–1146.

Schick, Anton (1986). "On asymptotically efficient estimation in semiparametric models". *Annals of Statistics* 14.3, pp. 1139–1151.

Schmidt-Hieber, Johannes (2020). "Nonparametric regression using deep neural networks with ReLU activation function". *Annals of Statistics* 48.4, pp. 1875–1897.

Semenova, Vira (2023). "Debiased machine learning of set-identified linear models". *Journal of Econometrics* 235.2, pp. 1725–1746.

Semenova, Vira and Victor Chernozhukov (2021). "Debiased machine learning of conditional average treatment effects and other causal functions". *Econometrics Journal* 24.2, pp. 264–289.

Semenova, Vira, Matt Goldman, Victor Chernozhukov, and Matt Taddy (2023). "Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence". *Quantitative Economics* 14.2, pp. 471–510.

Singh, Rahul and Liyang Sun (2024). "Double robustness for complier parameters and a semi-parametric test for complier characteristics". *Econometrics Journal* 27.1, pp. 1–20.

Sokolova, Anna and Todd Sorensen (2021). "Monopsony in Labor Markets: A Meta-Analysis". *ILR Review* 74.1, pp. 27–55.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225.2, pp. 175–199.

van der Vaart, Aad W. (1991). "On Differentiable Functionals". *Annals of Statistics* 19.1, pp. 178–204.

— (1998). *Asymptotic Statistics*. Cambridge University Press.

Varian, Hal R (2014). "Big data: New tricks for econometrics". *Journal of Economic Perspectives* 28.2, pp. 3–28.

Velez, Amilcar (2024). "On the Asymptotic Properties of Debiased Machine Learning Estimators". *arXiv:2411.01864*.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wiemann, Thomas (2026). "Optimal Categorical Instrumental Variables". *arXiv:2311.17021*.

Wiemann, Thomas, Achim Ahrens, Christian B Hansen, and Mark E Schaffer (2023). "`ddml`: Double/Debiased Machine Learning in R".

Wolpert, David H (1996). "The lack of a priori distinctions between learning algorithms". *Neural computation* 8.7, pp. 1341–1390.

Wright, Marvin N. and Andreas Ziegler (2017). "`ranger`: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". *Journal of Statistical Software* 77.1, pp. 1–17.

Wüthrich, Kaspar and Ying Zhu (2023). "Omitted variable bias of Lasso-based inference methods: A finite sample analysis". *Review of Economics and Statistics* 105.4, pp. 982–997.

Yatchew, Adonis (1998). "Nonparametric regression techniques in economics". *Journal of Economic Literature* 36.2, pp. 669–721.

Zhang, Cun-Hui and Stephanie S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 76.1, pp. 217–242.

Zheng, Mengchu, Matteo Bonvini, and Zijian Guo (2025). "Perturbed Double Machine Learning: Nonstandard Inference Beyond the Parametric Length". *arXiv:2511.01222*.

Zheng, Wenjing and Mark J. van der Laan (2011). "Cross-validated targeted minimum-loss-based estimation". In: *Targeted Learning*. Springer, pp. 459–474.