Electronic Journal of Statistics
Vol. 15 (2021) 6461–6543
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1931

Regularizing double machine learning in partially linear endogenous models*

Corinne Emmenegger and Peter Bühlmann

Seminar for Statistics,
ETH Zürich
e-mail: emmenegger@stat.math.ethz.ch; peter.buehlmann@stat.math.ethz.ch

Abstract: The linear coefficient in a partially linear model with confounding variables can be estimated using double machine learning (DML). However, this DML estimator has a two-stage least squares (TSLS) interpretation and may produce overly wide confidence intervals. To address this issue, we propose a regularization and selection scheme, *regsDML*, which leads to narrower confidence intervals. It selects either the TSLS DML estimator or a regularization-only estimator depending on whose estimated variance is smaller. The regularization-only estimator is tailored to have a low mean squared error. The regsDML estimator is fully data driven. The regsDML estimator converges at the parametric rate, is asymptotically Gaussian distributed, and asymptotically equivalent to the TSLS DML estimator, but regsDML exhibits substantially better finite sample properties. The regsDML estimator uses the idea of k-class estimators, and we show how DML and k-class estimation can be combined to estimate the linear coefficient in a partially linear endogenous model. Empirical examples demonstrate our methodological and theoretical developments. Software code for our regsDML method is available in the R-package `dmlalg`.

Keywords and phrases: Double machine learning, endogenous variables, generalized method of moments, instrumental variables, k-class estimation, partially linear model, regularization, semiparametric estimation, two-stage least squares.

Received January 2021.

1. Introduction

Partially linear models (PLMs) combine the flexibility of nonparametric approaches with ease of interpretation of linear models. Allowing for nonparametric terms makes the estimation procedure robust to some model misspecifications. A plaguing issue is potential endogeneity. For instance, if a treatment is not randomly assigned in a clinical study, subjects receiving different treatments differ in other ways than only the treatment [73]. Another situation where an explanatory variable is correlated with the error term occurs if the explanatory variable is determined simultaneously with the response [97]. In such situations, employing estimation methods that do not account for endogeneity can lead to biased estimators [44].

*This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461).

Let us consider the PLM

$$Y = X^T\beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y. \tag{1}$$

The covariates X and W and the response Y are observed whereas the variable H is not observed and acts as a potential confounder. It can cause endogeneity in the model when it is correlated with X, W, and Y. The variable ε_Y denotes a random error. An overview of PLMs is presented in Härdle, Liang and Gao [48]. Semiparametric methods are summarized in Ruppert, Wand and Carroll [83] and Härdle et al. [49], for instance.

Chernozhukov et al. [31] introduce double machine learning (DML) to estimate the linear coefficient β_0 in a model similar to (1). The central ingredients are Neyman orthogonality and sample splitting with cross-fitting. They allow estimates of so-called nuisance terms to be plugged into the estimating equation of β_0. The resulting estimator converges at the parametric rate $N^{-\frac{1}{2}}$, with N denoting the sample size, and is asymptotically Gaussian.

A common approach to cope with endogeneity uses instrumental variables (IVs). Consider a random variable A that typically satisfies the assumptions of a conditional instrument [76]. The DML procedure first adjusts A, X, and Y for W by regressing out W of them. Then the residual $Y - \mathbb{E}[Y|W]$ is regressed on $X - \mathbb{E}[X|W]$ using the instrument $A - \mathbb{E}[A|W]$. The population parameter is identified by

$$\beta_0 = \frac{\mathbb{E}\left[(A - \mathbb{E}[A|W])(Y - \mathbb{E}[Y|W])\right]}{\mathbb{E}\left[(A - \mathbb{E}[A|W])(X - \mathbb{E}[X|W])\right]} \tag{2}$$

if both A and X are 1-dimensional. The restriction to the 1-dimensional case is only for simplicity at this point. Below, we consider multivariate A and X. In practice, we insert potentially biased machine learning (ML) estimates of the nuisance parameters $\mathbb{E}[A|W]$, $\mathbb{E}[X|W]$, and $\mathbb{E}[Y|W]$ into this equation for β_0. Estimates of these nuisance parameters are typically biased if their complexity is regularized. Neyman orthogonal scores and sample splitting allow circumventing empirical process conditions to justify inserting ML estimators of nuisance parameters into estimating equations [19, 31].

Equation (2) has a two-stage least squares (TSLS) interpretation [91, 92, 16, 22, 13, 6]. As mentioned above, the residual term $Y - \mathbb{E}[Y|W]$ is regressed on $X - \mathbb{E}[X|W]$ using the instrument $A - \mathbb{E}[A|W]$. In entirely linear models, the following findings have been reported about TSLS and related procedures. The TSLS estimator has been observed to be highly variable, leading to overly wide confidence intervals. For instance, although ordinary least squares (OLS) is biased in the presence of endogeneity, it has been observed to be less variable [96, 71, 89, 34, 62]. The issue with large or nonexisting variance of TSLS (the order of existing moments of TSLS depends on the degree of overidentification [66, 67, 68]) is also coupled with the strength of the instrument [21, 86, 87, 35, 12].

Reducing the variability is sometimes possible by using k-class estimators [93, 51, 82, 54].

The k-class estimators have been developed for entirely linear models. The TSLS estimator is a k-class estimator with a fixed value of $k = 1$, and Anderson, Kunitomo and Morimune [8] recommend to not use fixed k-class estimators. Three particularly well-established k-class estimators are the limited information maximum likelihood (LIML) estimator [10, 4] and the Fuller(1) and Fuller(4) estimators [43]. They have been developed for entirely linear models to overcome some deficiencies of TSLS. If many instruments are present, LIML experiences some optimality properties [9]. Furthermore, the normal approximation for the finite sample estimator may be suboptimal for TSLS but useful for LIML [11, 7, 5]. However, LIML has no moments [67, 79, 80, 52]. The Fuller estimators overcome this problem. Having no moments can lead to poor squared error performance, especially in weak instrument situations [45]. On the other hand, the Fuller(1) estimator is approximately unbiased, and Fuller(4) has particularly low mean squared error (MSE) [43]. Takeuchi and Morimune [90] give further asymptotic optimality results of the Fuller estimators.

We propose a regularization-selection DML method using the idea of k-class estimators. We call our method *regsDML*. It is tailored to reduce variance and hence improve the MSE of the estimator of β_0. Nevertheless, regsDML converges at the parametric rate, and its coverage of confidence intervals for the linear coefficient β_0 remains (asymptotically) valid. Empirical simulations demonstrate that regsDML typically leads to shorter confidence intervals than LIML, Fuller(1), and Fuller(4), while it still attains the nominal coverage level.

1.1. Our contribution

Our contribution is twofold. First, we build on the work of Chernozhukov et al. [31] to estimate β_0 in the endogenous PLM (1) with multidimensional A and X such that its estimator $\hat{\beta}$ converges at the parametric rate, $N^{-\frac{1}{2}}$, and is asymptotically Gaussian. In contrast to Chernozhukov et al. [31], we formulate the underlying model as a structural equation model (SEM) and allow A and X to be multidimensional. We directly specify an identifiability condition of β_0 instead of giving additional conditional moment restrictions. The SEM may be overidentified in the sense that the dimension of A can exceed the dimension of X. Overidentification can lead to more efficient estimators [3, 18, 47] and more robust estimators [75]. Considering SEMs and an identifiability condition allows us to apply DML to more general situations than in Chernozhukov et al. [31].

Second, we propose a DML method that employs regularization and selection. This method is called regsDML, and we develop it in Section 4. It reduces the potentially excessive estimated standard deviation of DML because it selects either the TSLS DML estimator or a regularization-only estimator called regDML depending on whose estimated variance is smaller. The underlying idea of the regularization-only estimator regDML is similar to k-class estimation [93] and anchor regression [82, 23]. Both k-class estimation and anchor regression are designed for linear models and may require choosing a regularization parameter. Our approach is designed for PLMs, and the regularization parameter is data driven. Recently, Jakobsen and Peters [54] have proposed a related strategy for linear (structural equation) models; whereas they rely on testing for choosing the amount of regularization, we tailor our approach to reduce the MSE such that the coverage of confidence intervals for β_0 remains valid. The regsDML estimator converges at the parametric rate and is asymptotically Gaussian. In this sense, and in contrast to Jakobsen and Peters [54], regsDML focuses on statistical inference beyond point estimation with coverage guarantees not only in linear models but also in potentially complex partially linear ones. The regsDML estimator is asymptotically equivalent to the TSLS-type DML estimator, but regsDML may exhibit substantially better finite sample properties. Furthermore, our developments show how DML and k-class estimation can be combined to estimate the linear coefficient in an endogenous PLM.

Our approach allows flexible model specification. We only require that X enters linearly in (1) and that the other terms are additive. In particular, the form of the effect of W on A or of A on W is not constrained. This is partly similar to TSLS, which is robust to model misspecifications in its first stage because it does not rely on a correct specification of the instrument effect on the covariate [15]. The detailed assumptions on how the variables A, X, W, H, and Y interact are given in Section 2: the variable A needs to satisfy an assumption similar to that for a conditional instrument, but there is some flexibility.

We consider a motivating example to illustrate some of the points mentioned above. Figure 1 gives the SEM we generate data from and its associated causal graph [59, 74, 76, 77, 78, 64]. By convention, we omit error variables in a causal graph if they are mutually independent [76]. The variable A is similar to a conditional instrument given W.

$$
\begin{aligned}
&(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_4(\mathbf{0}, \mathbb{1})\\
&W \sim \pi \cdot \mathrm{Unif}([-1,1])\\
&A \leftarrow 3 \cdot \tanh(2W) + \varepsilon_A\\
&H \leftarrow 2 \cdot \sin(W) + \varepsilon_H\\
&X \leftarrow -|A| - 2 \cdot \tanh(W) - H + \varepsilon_X\\
&Y \leftarrow X + 0.5W^2 - 3 \cdot \cos(0.25\pi H) + \varepsilon_Y
\end{aligned}
$$

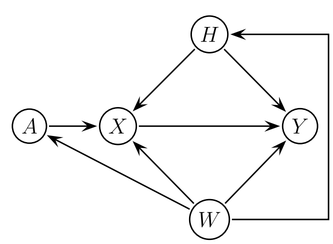

Fig 1. *An SEM and its associated causal graph.*

We simulate $M = 1000$ datasets each for a range of sample sizes N. The nuisance parameters $\mathbb{E}[A|W]$, $\mathbb{E}[X|W]$, and $\mathbb{E}[Y|W]$ are estimated with additive cubic B-splines with $\lceil N^{\frac{1}{5}} \rceil + 2$ degrees of freedom. The simulation results are displayed in Figure 2. This figure displays the coverage, power, and relative length of the 95% confidence intervals for β_0 using "standard" DML (red) and the newly proposed methods regDML (blue) and regsDML (green). The regDML

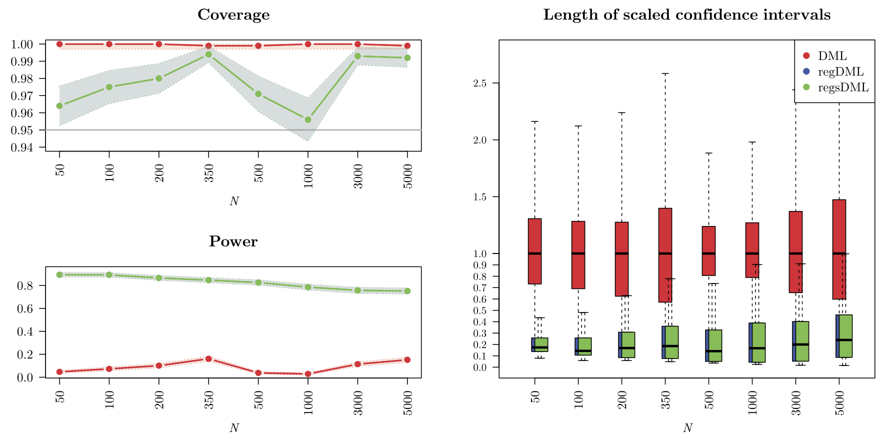

FIG 2. *The results come from* $M = 1000$ *simulation runs each from the SEM in Figure 1 for a range of sample sizes* N *and with* $K = 2$ *and* $\mathcal{S} = 100$ *in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for* β_0*, power for two-sided testing of the hypothesis* $H_0: \ \beta_0 = 0$*, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), and regsDML (green), where all results are at level* 95%*. At each* N*, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and power plots represent* 95% *confidence bands with respect to the* M *simulation runs. The blue and green lines are indistinguishable in the left panel.*

method is a version of regsDML with regularization only but no selection. If the blue curve is not visible in Figure 2, it coincides with the green curve. The dashed lines in the coverage and power plots indicate 95% confidence regions with respect to uncertainties in the M simulation runs.

The regsDML method succeeds in producing much narrower confidence intervals than DML although it maintains good coverage. The power of regsDML is close to 1 for all considered sample sizes. For small sample sizes, regsDML leads to confidence intervals whose length is around $10\% - 20\%$ the length of DML's. As the sample size increases, regsDML starts to resemble the behavior of the DML estimator but continues to produce substantially shorter confidence intervals. Thus, the regularization-selection regsDML (and also its version with regularization only) is a highly effective method to increase the power and sharpness of statistical inference whereas keeping the type I error and coverage under control.

Simulation results with $\beta_0 = 0$ in the SEM of Figure 2 are presented in Figure 7 in Section D in the appendix. Further numerical results are given in Section 5.

1.2. Additional literature

PLMs have received considerable interest. Härdle, Liang and Gao [48] present an overview of estimation methods in purely exogenous PLMs, and many references are given there. The remaining part of this paragraph refers to literature investigating endogenous PLMs. Ai and Chen [2] consider semiparametric estimation with a sieve estimator. Ma and Carroll [63] introduce a parametric model for the latent variable. Yao [98] considers a heteroskedastic error term and a partialling-out scheme [81, 85]. Florens, Johannes and Van Bellegem [42] propose to solve an ill-posed integral equation. Su and Zhang [88] investigate a partially linear dynamic panel data model with fixed effects and lagged variables and consider sieve IV estimators as well as an approach with solving integral equations. Horowitz [53] compares inference and other properties of nonparametric and parametric estimation if instruments are employed.

Combining Neyman orthogonality and sample splitting (with cross-fitting) allows a diverse range of estimators and machine learning algorithms to be used to estimate nuisance parameters. This procedure has alternatively been considered in Newey and McFadden [72], van der Laan and Robins [94], and Chernozhukov et al. [31]. DML methods have been applied in various situations. Chen, Huang and Tien [27] consider instrumental variables quantile regression. Liu, Zhang and Zhou [61] apply DML in logistic partially linear models. Colangelo and Lee [33] employ doubly debiased machine learning methods to a fully nonparametric equation of the response with a continuous treatment. Knaus [55] presents an overview of DML methods in unconfounded models. Farbmacher et al. [41] decompose the causal effect of a binary treatment by a mediation analysis and estimate it by DML. Lewis and Syrgkanis [60] extend DML to estimate dynamic effects of treatments. Chiang et al. [32] apply DML under multiway clustered sampling environments. Cui and Tchetgen Tchetgen [36] propose a technique to reduce the bias of DML estimators.

Nonparametric components can be estimated without sample splitting and cross-fitting if the underlying function class satisfies some entropy conditions; see for instance Mammen and van de Geer [65]. Alternatively, Chen, Liang and Zhou [28] partial out the nonparametric component using a kernel method and employ the generalized method of moments principle [46]. The mentioned entropy regularity conditions limit the complexity of the function class, and ML algorithms do usually not satisfy them. Particularly, these conditions fail to hold if the dimension of the nonparametric variables increases with the sample size [31].

Double robustness and orthogonality arguments have also been considered in the following works. Okui et al. [73] consider doubly robust estimation of the parametric part. Their estimator is consistent if either the model for the effect of the measured confounders on the outcome or the model of the effect of the measured confounders on the instrument is correctly specified. Smucler, Rotnitzky and Robins [84] consider doubly robust estimation of scalar parameters where the nuisance functions are ℓ_1-constrained. Targeted minimum loss based estimators and G-estimators also feature an orthogonality property; an overview is given in DiazOrdaz, Daniel and Kreif [38].

The literature presented in this subsection is related to but rather distinct from our work with the only exception of Chernozhukov et al. [31]. The difference to this latter contribution is highlighted in Section 2 and Section A in the appendix.

Outline of the paper. Section 2 and 3 describe the DML estimator. The former section introduces an identifiability condition, and the latter investigates asymptotic properties. Section 4 introduces the regularized regularization-selection estimator regsDML and its regularization-only version regDML and investigates their asymptotic properties. Section 5 presents numerical experiments and an empirical real data example. Section 6 concludes our work. Proofs and additional definitions and material are given in the appendix.

Notation. We denote by $[N]$ the set $\{1, 2, \ldots, N\}$. We add the probability law as a subscript to the probability operator $\mathbb{P}$ and the expectation operator $\mathbb{E}$ whenever we want to emphasize the corresponding dependence. We denote the $L^p(P)$ norm by $\|\cdot\|_{P,p}$ and the Euclidean or operator norm by $\|\cdot\|$, depending on the context. We implicitly assume that given expectations and conditional expectations exist. We denote by $\stackrel{d}{\to}$ convergence in distribution. Furthermore, we denote by $\mathbb{1}_{d\times d} \in \mathbb{R}^{d\times d}$ the $d \times d$ identity matrix and write $\mathbb{1}$ if we do not want to underline its dimension.

2. An identifiability condition and the DML estimator

Before we introduce regsDML in Section 4, we present our TSLS-type DML estimator of β_0 because we require it to formulate regsDML. The DML estimator estimates the linear coefficient in an endogenous and potentially overidentified PLM where A and X may be multidimensional. Our work builds on Chernozhukov et al. [31], but they only consider univariate A and X and restrict conditional moments to identify the linear coefficient. We impose an unconditional moment restriction below. However, our results recover theirs if A and X are univariate and the additional conditional moment restrictions are satisfied.

Our PLM is cast as an SEM. The SEM specifies the generating mechanism of the random variables A, W, H, X, and Y of dimensions q, v, r, d, and 1, respectively. The structural equation of the response is given by

$$Y \leftarrow X^T\beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y \tag{3}$$

as in (1), where $\beta_0 \in \mathbb{R}^d$ is a fixed unknown parameter vector, and where the functions g_Y and h_Y are unknown. The variable H is hidden and causes endogeneity. The variable ε_Y denotes an unobserved error term. The model is potentially overidentified in the sense that the dimension of A may exceed the dimension of X. Observe that A does not directly affect the response Y in the sense that it does not appear on the right hand side of (3). The model is required to satisfy an identifiability condition as in (5) below.

Econometric models are often presented as a system of simultaneous structural equations. Full information models consider all equations at once, and limited information models only consider equations of interest [5].

2.1. Identifiability condition

An identifiability condition is required to identify β_0 in (3). We define the residual terms

$$R_A := A - \mathbb{E}[A|W], \quad R_X := X - \mathbb{E}[X|W], \quad \text{and} \quad R_Y := Y - \mathbb{E}[Y|W] \tag{4}$$

that adjust A, X, and Y for W. Our DML estimator of β_0 is obtained by performing TSLS of R_Y on R_X using the instrument R_A. This scheme requires the unconditional moment condition

$$\mathbb{E}\left[R_A(R_Y - R_X^T\beta_0)\right] = \mathbf{0} \tag{5}$$

to identify β_0 in (3). For instance, this condition is satisfied if A is independent of both H and ε_Y given W or if A is independent of H, ε_Y, and W. The identifiability condition (5) is strictly weaker than the conditional moment conditions introduced in Chernozhukov et al. [31]; see Section A in the appendix that presents an example where our identifiability condition holds but the conditional moment conditions do not. The subsequent theorem asserts identifiability of β_0.

Theorem 2.1. *Let the dimensions $q = \dim(A)$ and $d = \dim(X)$, and assume $q \geq d$. Assume furthermore that the matrices $\mathbb{E}[R_X R_A^T]$ and $\mathbb{E}[R_A R_A^T]$ are of full rank, and assume the identifiability condition (5). Then the representation*

$$\beta_0 = \left(\mathbb{E}\left[R_X R_A^T\right]\mathbb{E}\left[R_A R_A^T\right]^{-1}\mathbb{E}\left[R_A R_X^T\right]\right)^{-1}\mathbb{E}\left[R_X R_A^T\right]\mathbb{E}\left[R_A R_A^T\right]^{-1}\mathbb{E}[R_A R_Y]$$

holds.

Theorem 2.1 precludes underidentification. The full rank condition of the matrix $\mathbb{E}[R_X R_A^T]$ expresses that the correlation between X and A is strong enough after regressing out W. This is a typical TSLS assumption [91, 92, 16, 22, 13, 6]. The rank assumptions in Theorem 2.1 in particular require that A, X, and Y are not deterministic functions of W.

The instrument A instead of R_A can alternatively identify β_0 in Theorem 2.1. However, this procedure leads to a suboptimal convergence rate of the resulting estimator; see Section 3.1.

The identifiability condition (5) is central to Theorem 2.1. Section G in the appendix presents examples illustrating SEMs where the identifiability condition holds and where it fails to hold.

2.2. Alternative interpretations of β_0

We present two alternative interpretations of β_0 apart from performing TSLS of R_Y on R_X using the instrument R_A. The second representation will be used to formulate our regularization schemes in Section 4. To formulate these alternative representations, we introduce the linear projection operator P_{R_A} on R_A that maps a random variable Z to its projection

$$P_{R_A} Z := \mathbb{E}\left[ZR_A^T\right]\mathbb{E}\left[R_AR_A^T\right]^{-1}R_A.$$

By Theorem 2.1, the population parameter β_0 solves the TSLS moment equation

$$\mathbf{0} = \mathbb{E}\left[R_XR_A^T\right]\mathbb{E}\left[R_AR_A^T\right]^{-1}\mathbb{E}\left[R_A(R_Y - R_X^T\beta_0)\right].$$

This motivates a generalized method of moments interpretation of β_0 because we have

$$\beta_0 = \underset{\beta\in\mathbb{R}^d}{\arg\min}\,\mathbb{E}[\psi(S;\beta,\eta^0)]\,\mathbb{E}\left[R_AR_A^T\right]^{-1}\mathbb{E}\left[\psi^T(S;\beta,\eta^0)\right]$$

for $\psi(S;\beta,\eta^0) := R_A(R_Y - R_X^T\beta)$, where $\eta^0 = (\mathbb{E}[A|W], \mathbb{E}[X|W], \mathbb{E}[Y|W])$ denotes the nuisance parameter and $S = (A, W, X, Y)$ denotes the concatenation of the observable variables.

This leads to the second interpretation of β_0. The coefficient β_0 minimizes the squared projection of the residual $R_Y - R_X^T\beta$ on R_A, namely

$$\beta_0 = \underset{\beta\in\mathbb{R}^d}{\arg\min}\,\mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T\beta)\right)^2\right]. \tag{6}$$

We employ the representation of β_0 in (6) to formulate our regularization schemes in Section 4.

3. Formulation of the DML estimator and its asymptotic properties

In this section, we describe how to estimate β_0 using the TSLS-type DML scheme, and we describe the asymptotic properties of this estimator.

Consider N iid realizations $\{S_i = (A_i, X_i, W_i, Y_i)\}_{i\in[N]}$ of $S = (A, X, W, Y)$ from the SEM in (3). We concatenate the observations of A row-wise to form an $(N\times q)$-dimensional matrix $\boldsymbol{A}$. Analogously, we construct the matrices $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ and $\boldsymbol{W} \in \mathbb{R}^{N\times v}$ and the vector $\boldsymbol{Y} \in \mathbb{R}^N$ containing the respective observations.

We construct a DML estimator of β_0 as follows. First, we split the data into $K \geq 2$ disjoint sets $I_1, \dots, I_K$. For simplicity, we assume that these sets are of equal cardinality $n = \frac{N}{K}$. In practice, their cardinality might differ due to rounding issues.

For each $k \in [K]$, we estimate the conditional expectations $m_A^0(W) := \mathbb{E}[A|W]$, $m_X^0(W) := \mathbb{E}[X|W]$, and $m_Y^0(W) := \mathbb{E}[Y|W]$, which act as nuisance parameters, with data from I_k^c. We call the resulting estimators $\hat{m}_A^{I_k^c}$, $\hat{m}_X^{I_k^c}$, and $\hat{m}_Y^{I_k^c}$, respectively. Then, the adjusted residual terms $\widehat{R}_{A,i}^{I_k} := A_i - \hat{m}_A^{I_k^c}(W_i)$, $\widehat{R}_{X,i}^{I_k} := X_i - \hat{m}_X^{I_k^c}(W_i)$, and $\widehat{R}_{Y,i}^{I_k} := Y_i - \hat{m}_Y^{I_k^c}(W_i)$ for $i \in I_k$ are evaluated on I_k, the complement of I_k^c. We concatenate them row-wise to form the matrices $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k} \in \mathbb{R}^{n\times q}$ and $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \in \mathbb{R}^{n\times d}$ and the vector $\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \in \mathbb{R}^n$.

These K iterates are assembled to form the DML estimator

$$\hat{\beta} := \left(\frac{1}{K}\sum_{k=1}^{K}\big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\right)^{-1} \frac{1}{K}\sum_{k=1}^{K}\big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \tag{7}$$

of β_0, where

$$\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} := \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\Big(\big(\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\big)^T \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\Big)^{-1}\big(\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\big)^T \tag{8}$$

denotes the orthogonal projection matrix onto the space spanned by the columns of $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}$.

To obtain $\hat{\beta}$ in (7), the individual matrices are first averaged before the final matrix is inverted. It is also possible to compute K individual TSLS estimators on the K iterates individually and average these. Both schemes are asymptotically equivalent. Chernozhukov et al. [31] call these two schemes DML2 and DML1, respectively, where DML2 is as in (7). The DML1 version of the coefficient estimator is given in the appendix in Section B.1. The advantage of DML2 over DML1 is that it enhances stability properties of the estimator. To ensure stability of the DML1 estimator, every individual matrix that is inverted needs to be well conditioned. Stability of the DML2 estimator is ensured if the average of these matrices is well conditioned.

The K sample splits are random. To reduce the effect of this randomness, we repeat the overall procedure $\mathcal{S}$ times and assemble the results as suggested in Chernozhukov et al. [31]. This procedure is described in Algorithm 1 in Section 4.2 below.

The following theorem establishes that $\hat{\beta}$ converges at the parametric rate and is asymptotically Gaussian.

Theorem 3.1. *Consider model* (3)*. Suppose that Assumption I.5 in the appendix in Section I holds and consider* $\overline{\psi}$ *given in Definition I.1 in the appendix in Section I. Then* $\hat{\beta}$ *as in* (7) *concentrates in a* $\frac{1}{\sqrt{N}}$ *neighborhood of* β_0*. It is approximately linear and centered Gaussian, namely*

$$\sqrt{N}\sigma^{-1}(\hat{\beta}-\beta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\psi}(S_i;\beta_0,\eta^0) + o_P(1) \xrightarrow{d} \mathcal{N}(0,\mathbb{1}_{d\times d}) \quad (N\to\infty),$$

uniformly over the law P *of* $S=(A,W,X,Y)$*, and where the variance-covariance matrix* σ^2 *is given by* $\sigma^2 = J_0\tilde{J}_0J_0^T$ *for the matrices* $\tilde{J}_0$ *and* J_0 *given in Definition I.1 in the appendix.*

A similar result to Theorem 3.1 is presented by Chernozhukov et al. [31]. However, their result requires univariate A and X, and it imposes conditional moment restrictions instead of the identifiability condition (5); see also Section A in the appendix that presents an example where our identifiability condition holds but the conditional moment conditions do not. If A and X are univariate and the respective conditional moment conditions hold, our result coincides with Chernozhukov et al. [31].

Theorem 3.1 also holds for the DML1 version of $\hat{\beta}$ defined in the appendix in Section B.1. Assumption I.5 specifies regularity conditions and the convergence rate of the machine learners estimating the conditional expectations. The machine learners are required to satisfy the product relations

$$\begin{aligned} &\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2}^2 \ll N^{-\frac{1}{2}}, \\ &\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2} \\ &\quad \cdot\big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \ll N^{-\frac{1}{2}} \end{aligned} \tag{9}$$

for $k \in [K]$, which allows us to employ a broad range of ML estimators. For instance, these convergence rates are satisfied by ℓ_1-penalized and related methods in a variety of sparse, high-dimensional linear models [26, 20, 24, 17], forward selection in sparse linear models [57], high-dimensional additive models [69, 56, 99], or regression trees and random forests [95, 14]. Please see Chernozhukov et al. [31] for additional references. In particular, the rate condition (9) is satisfied if the individual ML estimators converge at rate $N^{-\frac{1}{4}}$. Therefore, the individual ML estimators are not required to converge at rate $N^{-\frac{1}{2}}$.

The asymptotic variance σ^2 can be consistently estimated by replacing the true β_0 by $\hat{\beta}$ or its DML1 version. The nuisance functions are estimated on subsampled datasets, and the estimator of σ^2 is obtained by cross-fitting. The formal definition, the consistency result, and its proof are given in Definition I.1 and in Theorem I.21 in the appendix in Section I.

For fixed P, the asymptotic variance-covariance matrix σ^2 is the same as if the conditional expectations $m_A^0(W)$, $m_X^0(W)$, and $m_Y^0(W)$ and hence R_A, R_X, and R_Y were known.

Theorem 3.1 holds uniformly over laws P. This uniformity guarantees some robustness of the asymptotic statement [31]. The dimension v of the covariate W may grow as the sample size increases. Thus, high-dimensional methods can be considered to estimate the conditional expectations $\mathbb{E}[A|W]$, $\mathbb{E}[X|W]$, and $\mathbb{E}[Y|W]$.

The estimator $\hat{\beta}$ solves the moment equations

$$\mathbf{0} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i\in I_k}\widehat{R}_{X,i}^{I_k}\big(\widehat{R}_{A,i}^{I_k}\big)^T\Big(\frac{1}{n}\sum_{i\in I_k}\widehat{R}_{A,i}^{I_k}\big(\widehat{R}_{A,i}^{I_k}\big)^T\Big)^{-1}\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{\beta},\hat{\eta}^{I_k^c})\right),$$

where the score function ψ is given by

$$\psi(S;\beta,\eta) := \big(A - m_A(W)\big)\Big(Y - m_Y(W) - \big(X - m_X(W)\big)^T\beta\Big) \tag{10}$$

for $\eta = (m_A, m_X, m_Y)$, and where the estimated nuisance parameter is given by $\hat{\eta}^{I_k^c} = (\hat{m}_A^{I_k^c}, \hat{m}_X^{I_k^c}, \hat{m}_Y^{I_k^c})$. Observe that $\psi(S; \beta_0, \eta^0)$ with $\eta^0 = (m_A^0, m_X^0, m_Y^0)$ coincides with the term whose expectation is constrained to equal $\mathbf{0}$ in the identifiability condition (5). The crucial step to prove asymptotic normality of $\sqrt{N}(\hat{\beta} - \beta_0)$ is to analyze the asymptotic behavior of $\frac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(S_i; \hat{\beta}, \hat{\eta}^{I_k^c})$ for $k \in [K]$.

Apart from the identifiability condition, the first fundamental requirement to analyze these terms is the ML convergence rates in (9). Second, we employ sample splitting and cross-fitting. Sample splitting ensures that the data used to estimate the nuisance parameters and the data on which these estimators are evaluated are independent. Cross-fitting enables us to regain full efficiency. The third requirement is that the underlying score function ψ in (10) is Neyman orthogonal, which we explain next.

Neyman orthogonality ensures that ψ is insensitive to small changes in the nuisance parameter η at the true unknown linear coefficient β_0 and the true unknown nuisance parameter η^0. This makes estimation of β_0 robust to inserting biased ML estimators of the nuisance parameter in the estimation equation. The following definition formally introduces this concept.

Definition 3.2. *[31, Definition 2.1]. A score $\psi = \psi(S; \beta, \eta)$ is Neyman orthogonal at (β_0, η^0) if the pathwise derivative map*

$$\frac{\partial}{\partial r} \mathbb{E}\left[\psi\big(S; \beta_0, \eta^0 + r(\eta - \eta^0)\big)\right]$$

exists for all $r \in [0, 1)$ and nuisance parameters η and vanishes at $r = 0$.

Definition 3.2 does not entirely coincide with Chernozhukov et al. [31, Definition 2.1] because the latter also includes an identifiability condition. We directly assume the identifiability condition (5).

The subsequent proposition states that the score function ψ in (10) is indeed Neyman orthogonal.

Proposition 3.3. *The score ψ given in Equation (10) is Neyman orthogonal.*

We would like to remark that Neyman orthogonality of ψ neither depends on the distribution of S nor on the value of β_0 and η^0. In addition to being Neyman orthogonal, ψ is linear in β in the sense that we have

$$\psi(S; \beta, \eta) = \psi^b(S; \eta) - \psi^a(S; \eta)\beta \tag{11}$$

for

$$\psi^b(S; \eta) := \big(A - m_A(W)\big)\big(Y - m_Y(W)\big)$$

and

$$\psi^a(S; \eta) := \big(A - m_A(W)\big)\big(X - m_X(W)\big)^T.$$

This linearity property is also employed in the proof of Theorem 3.1.

3.1. Suboptimal estimation procedure

In general, we cannot employ A as an instrument instead of R_A in our TSLS-type DML estimation procedure. For simplicity, we assume $K = 2$ in this subsection and consider disjoint index sets I and I^c of size $n = \frac{N}{2}$. The term

$$\frac{1}{\sqrt{n}} \sum_{i \in I} A_i \big(\widehat{R}^I_{Y,i} - (\widehat{R}^I_{X,i})^T \beta_0\big) \tag{12}$$

may diverge as $N \to \infty$ because $\hat{m}^{I^c}_X$ and $\hat{m}^{I^c}_Y$ may be biased estimators of m^0_X and m^0_Y. This in particular happens if the functions m^0_X and m^0_Y are high-dimensional and need to be estimated by regularization techniques; see Chernozhukov et al. [31]. Even if sample splitting is employed, the term (12) is asymptotically not well behaved because the underlying score function

$$\varphi(S; \beta, \eta) := A\Big(Y - m_Y(W) - \big(X - m_X(W)\big)^T \beta\Big)$$

is not Neyman orthogonal. The issue is illustrated in Figure 3. The SEM used to generate the data is similar to the nonconfounded model used in Chernozhukov et al. [31, Figure 1]. The centered and rescaled term $\frac{\hat{\beta} - \beta_0}{\widehat{\mathrm{Var}}(\hat{\beta})}$ using A as an instrument is biased whereas it is not if the instrument R_A is used. Here, $\widehat{\mathrm{Var}}(\hat{\beta})$ denotes the empirically observed variance of $\hat{\beta}$ with respect to the performed simulation runs.

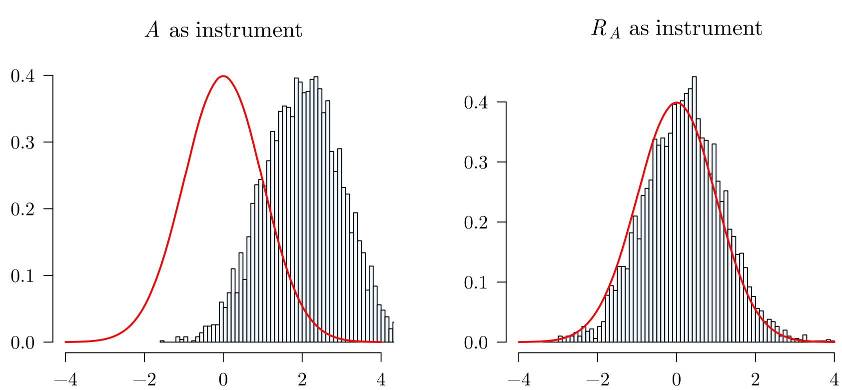

FIG 3. *Histograms of $\frac{\hat{\beta} - \beta_0}{\widehat{Var}(\hat{\beta})}$, where $\widehat{Var}(\hat{\beta})$ denotes the empirically observed variance of $\hat{\beta}$ with respect to the simulation runs, using A as an instrument in the left plot and using R_A as an instrument in the right plot. The orange curves represent the density of $\mathcal{N}(0, 1)$. The results come from* 5000 *simulation runs of sample size* 5000 *each from the SEM in the appendix in Section C with $K = 2$ and $\mathcal{S} = 1$. The conditional expectations are estimated with random forests consisting of* 500 *trees that have a minimal node size of* 5.

4. Regularizing the DML estimator: regDML and regsDML

We introduce a regularized estimator, regsDML, whose estimated standard deviation is typically smaller and never worse than the one of the TSLS-type DML estimator described above. Supporting theory and simulations illustrate that the associated confidence intervals nevertheless reach (asymptotically) valid and good coverage. The reg**s**DML estimator **s**elects either the DML estimator or its regularization-only version regDML, depending on which of the two has a smaller estimated standard deviation.

Subsequently, we first introduce the regularization-only method regDML. The regDML estimator is obtained by regularizing DML and choosing a data-dependent regularization parameter. Before we describe the choice of the regularization parameter, we introduce the regularization scheme for fixed regularization parameters.

Given a regularization parameter $\gamma \geq 0$, the population coefficient b^γ of the regularization scheme optimizes an objective function similar to the one used in k-class regression [93] or anchor regression [82, 23]. We established the representation

$$\beta_0 = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T\beta)\right)^2\right]$$

of β_0 in (6). For a regularization parameter $\gamma \geq 0$, we consider the regularized objective function and corresponding population coefficient

$$b^\gamma := \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\left((\mathrm{Id} - P_{R_A})(R_Y - R_X^T\beta)\right)^2\right] + \gamma\, \mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T\beta)\right)^2\right]. \quad (13)$$

This regularized objective is form-wise analogous to the objective function employed in anchor regression. The anchor regression estimator has been reformulated as a k-class estimator by Jakobsen and Peters [54] for a linear model.

If $\gamma = 1$, ordinary least squares regression of R_Y on R_X is performed. If $\gamma = 0$, we are partialling out or adjusting for the variable R_A. If $\gamma = \infty$, we perform TSLS regression of R_Y on R_X using the instrument R_A. In this case, b^γ coincides with β_0. The coefficient b^γ interpolates between the OLS coefficient $b^{\gamma=1}$ and the TSLS coefficient β_0 for general choices of $\gamma > 1$. For $\gamma > 1$, there is a one to one correspondence between b^γ and the k-class estimator (based on R_A, R_X, and R_Y) with regularization parameter $\kappa = \frac{\gamma-1}{\gamma} \in (0,1)$; see Jakobsen and Peters [54].

4.1. Estimation and asymptotic normality

In this section, we describe how to estimate b^γ in (13) for fixed $\gamma \geq 0$ using a DML scheme, and we describe the asymptotic properties of this estimator. We consider the residual matrices $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k} \in \mathbb{R}^{n\times q}$ and $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \in \mathbb{R}^{n\times d}$ and the vector

$\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \in \mathbb{R}^n$ introduced in Section 3 that adjust the data with respect to the nonparametric variables. The estimator of b^γ is given by

$$\hat{b}^\gamma := \arg\min_{b\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^K \Bigg(\Big\| \big(\mathbb{1} - \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\big(\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T b\big) \Big\|_2^2 \\ + \gamma \Big\| \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} (\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T b) \Big\|_2^2 \Bigg),$$

where $\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}$ is as in (8). This estimator can be expressed in closed form by

$$\hat{b}^\gamma = \left(\frac{1}{K}\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k}\big)^T \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k} \right)^{-1} \frac{1}{K}\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k}\big)^T \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k}, \tag{14}$$

where

$$\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k} := \Big(\mathbb{1} + (\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\Big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \quad \text{and} \quad \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k} := \Big(\mathbb{1} + (\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\Big)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}. \tag{15}$$

The computation of $\hat{b}^\gamma$ is similar to an OLS scheme where $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k}$ is regressed on $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k}$. To obtain $\hat{b}^\gamma$, individual matrices are first averaged before the final matrix is inverted. It is also possible to directly carry out the K OLS regressions of $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k}$ on $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k}$ and average the resulting parameters. Both schemes are asymptotically equivalent. We call the two schemes DML2 and DML1, respectively. This is analogous to Chernozhukov et al. [31] as already mentioned in Section 3. The DML1 version is presented in the appendix in Section B.2. As mentioned in Section 3, the advantage of DML2 over DML1 is that it enhances stability properties of the coefficient estimator because the average of matrices needs to be well conditioned but not every individual matrix.

Theorem 4.1. *Let $\gamma \geq 0$. Suppose that Assumption I.5 in the appendix in Section I (same as in Theorem 3.1) except I.5.1 holds, and consider the quantities $\sigma^2(\gamma)$ and $\overline{\psi}$ introduced in Definition J.1 in the appendix in Section J. The estimator $\hat{b}^\gamma$ concentrates in a $\frac{1}{\sqrt{N}}$ neighborhood of b^γ. It is approximately linear and centered Gaussian, namely*

$$\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^\gamma - b^\gamma) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \overline{\psi}(S_i; b^\gamma, \eta^0) + o_P(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d\times d}) \quad (N \to \infty),$$

uniformly over laws P of $S = (A, W, X, Y)$.

Theorem 4.1 also holds for the DML1 version of $\hat{b}^\gamma$ defined in the appendix in Section B.2. The influence function is denoted by $\overline{\psi}$ in both Theorems 3.1 and 4.1 but is defined differently. Assumption I.5 specifies regularity conditions and the convergence rate of the machine learners of the conditional expectations.

The machine learners are required to satisfy the product relations

$$\begin{aligned}
&\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2}^2 \ll N^{-\frac{1}{2}},\\
&\|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\\
&\quad\cdot\big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \ll N^{-\frac{1}{2}},\\
&\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2}\\
&\quad\cdot\big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \ll N^{-\frac{1}{2}}
\end{aligned}$$

for $k \in [K]$. The main difference to Theorem 3.1 and quantity of interest is the asymptotic variance $\sigma^2(\gamma)$. It can be consistently estimated with either $\hat{b}^\gamma$ or its DML1 version as illustrated in Theorem J.3 in the appendix in Section J. Typically, for $\gamma < \infty$, the asymptotic variance $\sigma^2(\gamma)$ is smaller than σ^2 in Theorem 3.1. Such a variance gain comes at the price of bias because $\hat{b}^\gamma$ estimates b^γ and not the true parameter β_0.

The proof of Theorem 4.1 uses Neyman orthogonality of the underlying score function. Recall that Neyman orthogonality neither depends on the distribution of S nor on the value of the coefficients β_0 and η^0 as discussed in Section 3.

For fixed $\gamma > 1$, Theorem 4.1 furthermore implies that the k-class estimator corresponding to $\hat{b}^\gamma$ converges at the parametric rate and follows a Gaussian distribution asymptotically.

4.2. Estimating the regularization parameter γ

For simplicity, we assume $d = 1$ in this subsection. The results can be extended to $d > 1$.

Subsequently, we introduce a data-driven method to choose the regularization parameter γ in practice. The estimated regularization parameter γ leads to an estimate of β_0 that asymptotically has the same MSE behavior as the TSLS-type estimator $\hat{\beta}$ in (7) but may exhibit substantially better finite sample properties.

We consider the estimated regularization parameter

$$\hat{\gamma} := \arg\min_{\gamma \geq 0} \frac{1}{N}\hat{\sigma}^2(\gamma) + |\hat{b}^\gamma - \hat{\beta}|^2. \tag{16}$$

It optimizes an estimate of the asymptotic MSE of $\hat{b}^\gamma$: the term $\hat{\sigma}^2(\gamma)$ is the consistent estimator of $\sigma^2(\gamma)$ described in Theorem J.3 in the appendix in Section J, and the term $|\hat{b}^\gamma - \hat{\beta}|^2$ is a plug-in estimator of the squared population bias $|b^\gamma - \beta_0|^2$. The estimated regularization parameter $\hat{\gamma}$ is random because it depends on the data.

First, we investigate the bias of the population parameter b^{γ_N} for a nonrandom sequence of regularization parameters $\{\gamma_N\}_{N\geq 1}$ as $N \to \infty$. Afterwards, we propose a modified estimator of the regularization parameter whose corresponding parameter estimate is denoted by regDML, and we introduce the regularization-selection estimator regsDML. Finally, we analyze the asymptotic properties of regDML and regsDML.

Let us consider a deterministic sequence $\{\gamma_N\}_{N\geq 1}$ of regularization parameters. By Proposition 4.2 below, the (scaled) population bias $\sqrt{N}|b^{\gamma_N}-\beta_0|$ vanishes as $N\to\infty$ if γ_N is of larger order than $\sqrt{N}$.

Proposition 4.2. *Suppose that I.5.1, I.5.3, and I.5.4 of Assumption I.5 in the appendix in Section I hold (subset of the assumptions in Theorem 3.1). Assume $\{\gamma_N\}_{N\geq 1}$ is a sequence of non-negative real numbers. Then we have*

$$\sqrt{N}|b^{\gamma_N}-\beta_0| \to \begin{cases} 0, & \text{if } \gamma_N \gg \sqrt{N} \\ C, & \text{if } \gamma_N \sim \sqrt{N} \\ \infty, & \text{if } \gamma_N \ll \sqrt{N} \end{cases}$$

as $N\to\infty$ for some non-negative finite real number C.

Theorem 4.3 below shows that the estimated regularization parameter $\hat\gamma$ is of equal or larger stochastic order than $\sqrt{N}$. If it were not, choosing $\gamma=\infty$ in (16), and hence selecting the TSLS-type estimator $\hat\beta$, would lead to a smaller estimated asymptotic MSE.

Theorem 4.3. *Let $\gamma_N = o(\sqrt{N})$, and suppose that Assumption I.5 in the appendix in Section I holds (same as in Theorem 3.1). We then have*

$$\lim_{N\to\infty} P\big(\hat\sigma^2(\gamma_N) + N(\hat b^{\gamma_N}-\hat\beta)^2 \leq \hat\sigma^2\big) = 0.$$

If $\hat\gamma$ is multiplied by a deterministic scalar a_N that diverges to $+\infty$ at an arbitrarily slow rate as $N\to\infty$, the modified regularization parameter $\hat\gamma' := a_N\hat\gamma$ is of stochastic order larger than $\sqrt{N}$. By default, we choose $a_N=\log(\sqrt{N})$. Proposition 4.2 is formulated for deterministic regularization parameters, but the deterministic statements can be replaced by probabilistic ones. Proposition 4.2 then implies that the population bias term $|b^{\hat\gamma'}-\beta_0|$ vanishes at rate $o_P(N^{-\frac{1}{2}})$. Thus, the two quantities $\sqrt{N}(\hat b^{\hat\gamma'}-b^{\hat\gamma'})$ and $\sqrt{N}(\hat b^{\hat\gamma'}-\beta_0)$ are asymptotically equivalent due to Theorem 4.4 below, and we have

$$\sqrt{N}(\hat b^{\hat\gamma'}-\beta_0) \approx \mathcal{N}\big(0,\sigma^2(\hat\gamma')\big)$$

whenever N is sufficiently large (note that asymptotically as $N\to\infty$, the right-hand side has the same limit as described in Theorem 4.4).

We call $\hat b^{\hat\gamma'}$ the regDML (regularized DML) estimator. The regularization-selection estimator selects between DML and regDML based on whose variance estimate is smaller. The "s" in regsDML stands for selection.

Theorem 4.4. *Suppose that Assumption I.5 in the appendix in Section I holds (same as in Theorem 3.1). Let $\{a_j\}_{j\geq 1}$ be a sequence of deterministic, non-negative real numbers that diverges to ∞ as $N\to\infty$. Furthermore, consider*

$\hat{\gamma}' = a_N\hat{\gamma}$ as above. Then, we have

$$\sqrt{N}\hat{\sigma}^{-1}(\hat{\gamma}')(\hat{b}^{\hat{\gamma}'} - b^{\hat{\gamma}'}) = \sqrt{N}\sigma^{-1}(\hat{\beta} - \beta_0) + o_P(1)$$

uniformly over laws P of $S = (A, W, X, Y)$, where $\hat{\sigma}(\cdot)$ ist the estimator from Theorem J.3 in the appendix, which consistently estimates $\sigma(\cdot)$ from 4.1.

Particularly, $\hat{b}^{\hat{\gamma}'}$ and $\hat{\beta}$ are asymptotically equivalent. But $\hat{b}^{\hat{\gamma}'}$ may exhibit substantially better finite sample properties as we demonstrate in the subsequent section. Because $\hat{b}^{\hat{\gamma}'}$ and $\hat{\beta}$ are asymptotically equivalent, the same result also holds for the selection estimator regsDML.

The proof of Theorem 4.4 does not depend on the precise construction of $\hat{\gamma}'$ and only uses that the random regularization parameter is of stochastic order larger than $\sqrt{N}$. Thus, Theorem 4.4 remains valid if the regularization parameter comes from k-class estimation and is of the required stochastic order. The same stochastic order is also required to show that k-class estimators are asymptotically Gaussian [70, 68].

The K sample splits are random. To reduce the effect of this randomness, we repeat the overall procedure $\mathcal{S}$ times and assemble the results as suggested in Chernozhukov et al. [31]. The assembled parameter estimate is given by the median of the individual parameter estimates; see Steps 9 and 10 of Algorithm 1. The assembled variance estimate is given by adding a correction term to the individual variances and subsequently taking the median of these corrected terms. The correction term measures the variability due to sample spitting across $s \in [\mathcal{S}]$.

It is possible that the assembled variance of regDML is larger than the assembled variance of DML. In such a case, we do not use the regDML estimator and select the DML estimator instead to ensure that the final estimator of β_0 does not experience a larger estimated variance than DML. This is the regsDML scheme. A summary of this procedure is given in Algorithm 1.

5. Numerical experiments

This section illustrates the performance of the DML, regDML, and regsDML estimators in a simulation study and for an empirical dataset. Our implementation is available in the R-package `dmlalg` [40]. We employ the DML2 method and $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. Furthermore, we compare our estimation schemes with the following three k-class estimators: LIML, Fuller(1), and Fuller(4). On each of the K sample splits, we compute the regularization parameter of the respective k-class estimation procedure and average them. Then, we compute the corresponding γ-value and proceed as for the other regularized estimators according to Algorithm 1.

The first example in Section 5.1 considers an overidentified model in which the dimension of A is larger than the dimension of X. The second example in

Algorithm 1: regsDML in a PLM with confounding variables.

Input : N iid realizations from the SEM (3), a natural number $\mathcal{S}$, a regularization parameter grid $\{\gamma_i\}_{i\in[L]}$ for some natural number L, a non-negative diverging sequence $\{a_j\}_{j\geq 1}$.
Output: An estimator of β_0 in (3) together with its estimated asymptotic variance.

for $s \in [\mathcal{S}]$ **do**
  Compute $\hat{\beta}_s = \hat{\beta}$ and $\hat{\sigma}_s^2 = \hat{\sigma}^2$.
  Compute $\hat{b}_s^{\gamma_i} = \hat{b}^{\gamma_i}$ and $\hat{\sigma}_s^2(\gamma_i) = \hat{\sigma}^2(\gamma_i)$ for $i \in [L]$.
  Choose $\hat{\gamma}_s = \arg\min_{\gamma\in\{\gamma_i\}_{i\in[L]}} \big(\frac{1}{N}\hat{\sigma}_s^2(\gamma) + |\hat{b}_s^{\gamma} - \hat{\beta}_s|^2\big)$ and let $\hat{\gamma}_s' = a_N\hat{\gamma}_s$.
  Compute $\hat{b}_s^{\hat{\gamma}_s'} = \hat{b}^{\hat{\gamma}_s'}$ and $\hat{\sigma}_s^2(\hat{\gamma}_s') = \hat{\sigma}^2(\hat{\gamma}_s')$.
end
Compute $\hat{\beta}^{\mathrm{med}} = \mathrm{median}_{s\in[\mathcal{S}]}(\hat{\beta}_s)$.
Compute $\hat{b}_{\mathrm{reg}}^{\mathrm{med}} = \mathrm{median}_{s\in[\mathcal{S}]}(\hat{b}_s^{\hat{\gamma}_s'})$.
Compute $\hat{\sigma}^{2,\mathrm{med}} = \mathrm{median}_{s\in[\mathcal{S}]}\big(\hat{\sigma}_s^2 + (\hat{\beta}_s - \hat{\beta}^{\mathrm{med}})^2\big)$.
Compute $\hat{\sigma}_{\mathrm{reg}}^{2,\mathrm{med}} = \mathrm{median}_{s\in[\mathcal{S}]}\big(\hat{\sigma}_s^2(\hat{\gamma}_s') + (\hat{b}_s^{\hat{\gamma}_s'} - \hat{b}_{\mathrm{reg}}^{\mathrm{med}})^2\big)$.
if $\hat{\sigma}_{\mathrm{reg}}^{2,\mathrm{med}} < \hat{\sigma}^{2,\mathrm{med}}$ **then**
  Take the parameter estimate $\hat{b}_{\mathrm{reg}}^{\mathrm{med}}$ together with its associated estimated asymptotic variance $\frac{1}{N}\hat{\sigma}_{\mathrm{reg}}^{2,\mathrm{med}}$.
else
  Take the parameter estimate $\hat{\beta}^{\mathrm{med}}$ together with its associated estimated asymptotic variance $\frac{1}{N}\hat{\sigma}^{2,\mathrm{med}}$.
end

Section 5.2 considers justidentified real-world data. In both examples, the conditional expectations acting as nuisance parameters are estimated with random forests.

An example where the conditional expectations are estimated with splines is given in Section 1.1. Additional empirical results are provided in the appendix in Section D, E, and F. The latter section considers examples where DML, regDML, and regsDML do not work well in finite sample situations: we follow the NCP (No Cherry Picking) guideline [25] to possibly enhance further insights into the finite sample behavior. Section E in the appendix presents examples where the link $A \to X$ is weak and examples illustrating the bias-variance tradeoff of the respective estimated quantities as a function of the regularization parameter γ.

5.1. Simulation example with random forests

We generate data from the SEM in Figure 4. This SEM satisfies the identifiability condition (5) because A_1 and A_2 are independent of H given W_1 and W_2; a proof is given in the appendix in Section K. The model is overidentified because the dimension of $A = (A_1, A_2)$ is larger than the dimension of X. The variable A_1 directly influences A_2 that in turn directly affects W_1. Both W_1 and W_2 directly influence H. Both A_1 and A_2 directly influence X. The variable A_1 is a source node.

$$
\begin{aligned}
&(\varepsilon_{A_1}, \varepsilon_{A_2}, \varepsilon_{W_1}, \varepsilon_{W_2}, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_7(\mathbf{0}, \mathbb{1})\\
&A_1 \leftarrow \mathbb{1}_{\{\varepsilon_{A_1} \leq 0\}}\\
&A_2 \leftarrow -4A_1 + \varepsilon_{A_2}\\
&W_1 \leftarrow 2A_2 + \varepsilon_{W_1}\\
&W_2 \leftarrow \varepsilon_{W_2}\\
&H \leftarrow 2\mathbb{1}_{\{\sin(\pi W_1)\cdot\tanh(W_2)\geq 0\}} + \varepsilon_H\\
&X \leftarrow 1.5A_1 - 0.5A_2 + \tanh(H)\\
&\qquad -2\mathbb{1}_{\{W_1 \geq 0\}}\mathbb{1}_{\{W_2 \leq 0\}} + \varepsilon_X\\
&Y \leftarrow X + \mathbb{1}_{\{W_2 \leq 0\}} + \sin(\pi H) + \varepsilon_Y
\end{aligned}
$$

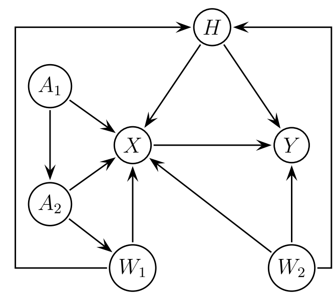

Fig 4. *An SEM and its associated causal graph.*

We simulate $M = 1000$ datasets each from the SEM in Figure 4 for a range of sample sizes. For every dataset, we compute a parameter estimate and an associated confidence interval with DML, regDML, and regsDML. We choose $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1 and estimate the conditional expectations with random forests consisting of 500 trees that have a minimal node size of 5.

Figure 5 illustrates our findings. It gives the coverage, power, and relative length of the 95% confidence intervals for a range of sample sizes N of the three methods. The blue and green curves correspond to regDML and regsDML, respectively. If the blue curve is not visible in Figure 5, it coincides with the green one. The two regularization methods perform similarly because regularization can considerably improve DML. The red curves correspond to DML. If the red curve is not visible, it coincides with LIML, whose results are displayed in orange. The Fuller(1) and Fuller(4) estimators correspond to purple and cyan, respectively.

The top left plot in Figure 5 displays the coverages as interconnected dots. The dashed lines represent 95% confidence regions of the coverages. These confidence regions are computed with respect to uncertainties in the M simulation runs. No coverage region falls below the nominal 95% level that is marked by the gray line.

The bottom left plot in Figure 5 shows that the power of DML, LIML, and Fuller(1) is lower for small sample sizes and increases gradually. The power of the other regularization methods remains approximately 1. The dashed lines represent 95% confidence regions that are computed with respect to uncertainties in the M simulation runs.

The right plot in Figure 5 displays boxplots of the scaled lengths of the confidence intervals. For each N, the confidence interval lengths of all methods are divided by the median confidence interval lengths of DML. The length of the regsDML confidence intervals is around $50\% - 80\%$ the length of DML's. Nevertheless, the coverage of regsDML remains around 95%. The LIML, Fuller(1), and Fuller(4) confidence intervals are considerably longer than regsDML's. Although the confidence intervals of regsDML are the shortest of all considered methods, its coverage remains valid.

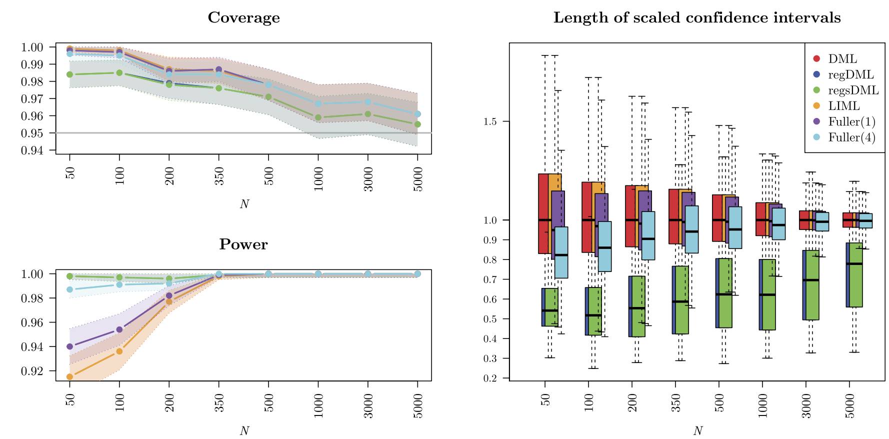

FIG 5. *The results come from $M = 1000$ simulation runs each from the SEM in Figure 4 for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with random forests. The figure displays the coverage of two-sided confidence intervals for β_0, power for two-sided testing of the hypothesis $H_0:\ \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the power plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines as well as the red and orange ones are indistinguishable in the left panel.*

Simulation results with $\beta_0 = 0$ in the SEM in Figure 4 are presented in Figure 8 in the appendix in Section D.

5.2. Real data example

We apply the DML and regsDML methods to a real dataset. We estimate the linear effect β_0 of institutions on economic performance following the work of Acemoglu, Johnson and Robinson [1] and Chernozhukov et al. [31]. Countries with better institutions achieve a greater level of income per capita, and wealthy economies can afford better institutions. This may cause simultaneity. To overcome it, mortality rates of the first European settlers in colonies are considered as a source of exogenous variation in institutions. For further details, we refer to Acemoglu, Johnson and Robinson [1] and Chernozhukov et al. [31]. The data is available in the R-package hdm [29] and is called AJR. In our notation, the response Y is the GDP, the covariate X the average protection against expropriation risk, the variable A the logarithm of settler mortality, and the covariate W consists of the latitude, the squared latitude, and the binary factors Africa, Asia, North America, and South America. That is, we adjust nonparametrically for the latitude and geographic information.

Table 1
Coefficient estimate, its standard error, and a confidence interval with DML and regsDML on the AJR dataset, where $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1, and where the conditional expectations are estimated with random forests consisting of 1000 trees that have a minimal node size of 5.

	Estimate of β_0	Standard error	Confidence interval for β_0
DML	0.739	0.459	$[-0.161, 1.639]$
regsDML	0.688	0.229	$[0.239, 1.136]$

We choose $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1 and compute the conditional expectations with random forests with 1000 trees that have a minimal node size of 5. The estimation results are displayed in Table 1. This table gives the estimated linear coefficient, its standard deviation, and a confidence interval for β_0 for DML and regsDML. The coefficient estimate of DML is not significant because the respective confidence interval includes 0. The regsDML estimate is significant because it has a smaller standard deviation than the DML estimate. Note that the coefficient estimate of regsDML falls within the DML confidence interval.

The AJR dataset has also been analyzed in Chernozhukov et al. [31]. They also estimate conditional expectations with random forests consisting of 1000 trees that have a minimal node size of 5 but implicitly assume an additional homoscedasticity condition for the errors $R_Y - R_X^T\beta_0$; see Chernozhukov et al. [30]. Such a homoscedastic error assumption is questionable though. Their procedure leads to a smaller estimate of the standard deviation of DML than what we obtain.

6. Conclusion

We extended and regularized double machine learning (DML) in potentially overidentified partially linear models (PLMs) with hidden variables. Our goal was to estimate the linear coefficient β_0 of the PLM. Hidden variables confound the observables, which can cause endogeneity. For instance, a clinical study may experience an endogeneity issue if a treatment is not randomly assigned and subjects receiving different treatments differ in other ways than the treatment [73]. In such situations, employing estimation methods that do not account for endogeneity leads to biased estimators [44].

Our contribution was twofold. First, we formulated the PLM as a structural equation model (SEM) and imposed an identifiability condition on it to recover the population parameter β_0. We estimated β_0 using DML similarly to Chernozhukov et al. [31]. However, our setting is more general than the one considered in Chernozhukov et al. [31] because we allow the predictors to be multivariate, and we impose a moment condition instead of restricting conditional moments. The DML estimation procedure allows biased estimators of additional nuisance functions to be plugged into the estimating equation of β_0. The resulting estimator of β_0 is asymptotically Gaussian and converges at the parametric rate of $N^{-\frac{1}{2}}$. However, DML has a two-stage least squares (TSLS) interpretation and may therefore lead to overly wide confidence intervals.

Second, we proposed a regularization-only DML scheme, regDML, and a regularization-selection DML scheme, regsDML. The latter has shorter confidence intervals by construction because it selects between DML and regDML depending on whose estimated standard deviation is smaller. Although regsDML and plain DML are asymptotically equivalent, regsDML leads to drastically shorter confidence intervals for finite sample sizes. Nevertheless, coverage guarantees for β_0 remain. The regDML estimator is similar to k-class estimation [93] and anchor regression [82, 23, 54] but allows potentially complex partially linear models and chooses a data-driven regularization parameter.

Empirical examples demonstrated our methodological and theoretical developments. The results showed that regsDML is a highly effective method to increase the power and sharpness of statistical inference. The DML estimator has a TSLS interpretation. Therefore, if the confounding is strong, the DML estimator leads to overly wide confidence intervals and can be substantially biased. In such a case, regsDML drastically reduces the width of the confidence intervals but may inherit additional bias from DML. This effect can be particularly pronounced for small sample sizes. Section F in the appendix presents examples with strong and reduced confounding and demonstrates the coverage behavior of DML and regsDML. Section E in the appendix analyzes the performance of our methods if the strength of the link $A \to X$ varies and investigates the bias-variance tradeoff of the respective estimated quantities for different values of the regularization parameter.

Although a wide range of machine learners can be employed to estimate the nuisance functions, we observed that additive splines can estimate more precise results than random forests if the underlying structure is additive in good approximation. This effect is particularly pronounced if the sample size is small. If such a finding is to be expected, it may be worthwhile to use structured models rather than "general" machine learning algorithms, especially with small or moderate sample size. Our regsDML methodology can be used with the implementation that is available in the R-package `dmlalg` [40].

Appendix A: An example where the identifiability condition (5) holds, but conditional moment requirements do not

This section presents an SEM where our identifiability condition (5) holds, but where the conditional moment requirements of Chernozhukov et al. [31] do not.

Let $d = 1 = q$ in this section (justidentified case), and assume the model

$$Y \leftarrow X\beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y$$

given in (3) and the identifiability condition $\mathbb{E}[R_A(R_Y - R_X\beta_0)] = 0$ given in (5).

Chernozhukov et al. [31] assume the model

$$Y = X\beta_0 + g_Y(W) + U, \quad A = g_A(W) + V \tag{17}$$

for unknown functions g_Y and g_A and impose the conditional moment restrictions

$$\mathbb{E}[U|A, W] = 0 \quad \text{and} \quad \mathbb{E}[V|W] = 0 \tag{18}$$

on the error terms.

Model (17) and the conditional moment restrictions (18) imply the identifiability condition (5) due to

$$\mathbb{E}\left[R_A(R_Y - R_X\beta_0)\right] = \mathbb{E}\left[\big(A - g_A(W)\big)U\right] = \mathbb{E}\left[\big(A - g_A(W)\big)\,\mathbb{E}[U|A, W]\right] = 0.$$

However, the reverse direction does not hold. A counterexample is presented in Figure 6 where W directly affects H. This SEM satisfies the identifiability condition (5) because A is independent of H conditional on W, but it does not satisfy $\mathbb{E}[U|W, A] = 0$ because we have

$$\mathbb{E}[U|A, W] = \mathbb{E}[H + \varepsilon_Y|A, W] = \mathbb{E}[H|W] = \mathbb{E}[W + \varepsilon_H|W] = W$$

due to $A \perp\!\!\!\perp H|W$ and $(\varepsilon_Y, \varepsilon_H) \perp\!\!\!\perp (W, A)$. We have $A \perp\!\!\!\perp H|W$ because all paths from A to H are blocked by W. The path $A \to X \leftarrow H$ is blocked by the empty set because X is a collider on this path. The path $A \to X \to Y \leftarrow H$ is blocked by the empty set because Y is a collider on this path. The path $A \to X \to Y \leftarrow W \to H$ is blocked by W. The paths $A \to X \to W \to Y \leftarrow H$ and $A \to X \to W \to H$ are also blocked by W.

$$\begin{aligned}
&(\varepsilon_A, \varepsilon_W, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_5(\mathbf{0}, \mathbb{1})\\
&A \leftarrow \varepsilon_A\\
&W \leftarrow \varepsilon_W\\
&H \leftarrow W + \varepsilon_H\\
&X \leftarrow A + W + H + \varepsilon_X\\
&Y \leftarrow X + W + H + \varepsilon_Y
\end{aligned}$$

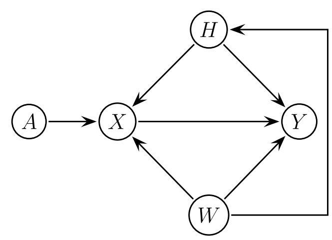

Fig 6. *An SEM and its associated causal graph.*

Appendix B: DML1 estimators

The DML1 estimators are less preferred than the DML2 estimators we proposed to use in the main text, but for completeness we provide the definitions in this section.

B.1. DML1 estimator of β_0

The DML1 estimator of β_0 is given by

$$\hat{\beta}^{\mathrm{DML1}} := \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}^{I_k},$$

where

$$\hat{\beta}^{I_k} := \Big((\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \Big)^{-1} (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}, \tag{19}$$

and where we recall the projection matrix $\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} = \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\big((\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k})^T\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\big)^{-1}(\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k})^T$ defined in (8). The estimator $\hat{\beta}^{I_k}$ is the TSLS estimator of $\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}$ on $\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}$ using the instrument $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}$.

B.2. DML1 estimator of b^γ

The DML1 estimator of b^γ is given by

$$\hat{b}^{\gamma,\mathrm{DML1}} := \frac{1}{K}\sum_{k=1}^{K}\hat{b}_k^{\gamma}, \tag{20}$$

where

$$\hat{b}_k^{\gamma} := \operatorname*{arg\,min}_{b\in\mathbb{R}^d}\bigg(\Big\| \big(\mathbb{1} - \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\big(\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T b\big) \Big\|_2^2 + \gamma \Big\| \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T b\big) \Big\|_2^2 \bigg).$$

This estimator can be expressed in closed form by

$$\hat{b}_k^{\gamma} = \Big((\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k})^T \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k} \Big)^{-1} (\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k})^T \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k},$$

where we recall the notation

$$\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k} = \Big(\mathbb{1} + (\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\Big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \quad \text{and} \quad \widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k} = \Big(\mathbb{1} + (\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\Big)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}$$

as in (15). The computation of $\hat{b}_k^{\gamma}$ is an OLS scheme where $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{Y}}}^{I_k}$ is regressed on $\widehat{\boldsymbol{R}}_{\tilde{\boldsymbol{X}}}^{I_k}$.

Appendix C: SEM of Figure 3

The data from the simulation displayed in Figure 3 come from the following SEM. Let the dimension of W be $v = 20$. Let R be the upper triangular matrix of the Cholesky decomposition of the Toeplitz matrix whose first row is given by $(1, 0.7, 0.7^2, \ldots, 0.7^{19})$. The SEM we consider is given by

$$
\begin{aligned}
&(\varepsilon_A, \varepsilon_W, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_{24}(\mathbf{0}, \mathbb{1}) \\
&H \leftarrow \varepsilon_H \\
&W \leftarrow \varepsilon_W R \\
&A \leftarrow \tfrac{e^{W_1}}{1+e^{W_1}} + W_2 + W_3 + \varepsilon_A \\
&X \leftarrow 2A + W_1 + 0.25 \cdot \tfrac{e^{W_3}}{1+e^{W_3}} + H + \varepsilon_X \\
&Y \leftarrow X + \tfrac{e^{W_1}}{1+e^{W_1}} + 0.25 W_3 + H + \varepsilon_Y.
\end{aligned}
$$

Appendix D: Additional numerical results

If we say in this section that the nuisance parameters are estimated with additive splines, they are estimated with additive cubic B-splines with $\lceil N^{\frac{1}{5}} \rceil + 2$ degrees of freedom, where N denotes the sample size of the data. If we say in this section that the nuisance parameters are estimated with random forests, they are estimated with random forests consisting of 500 trees that have a minimal node size of 5.

Figure 7 and 8 illustrate the simulation results with $\beta_0 = 0$ of the examples presented in Figure 2 and 5 in Section 1.1 and 5.1, respectively. The coverage and length of the scaled confidence intervals are similar to the results obtained for $\beta_0 \neq 0$. Instead of the power as in Figure 2 and 5, Figure 7 and 8 illustrate the type I error.

In Figure 7, DML achieves a type I error of 0 or close to 0 over all sample sizes considered. The regsDML method achieves a type I error that is closer to the gray line indicating the 5% level. The dashed lines represent 95% confidence regions. The type I error of regsDML is higher than the type I error of DML because the regsDML confidence intervals are considerably shorter than the DML ones. The right plot in Figure 7 indicates that the length of the confidence intervals of regsDML is around $10\% - 30\%$ the length of DML's. Although regsDML greatly reduces the confidence interval length, the type I error confidence bands include the 5% level or are below it. This means that although regsDML is a regularized version of DML, it does not incur an overlarge bias.

In Figure 8, the type I errors of both DML and regsDML are similar. The 95% confidence regions of both estimators, which are represented by dashed lines, include the 5% level or are below it. The right plot in Figure 8 illustrates that the regsDML confidence intervals are around $50\% - 80\%$ the length of DML's. Nevertheless, its type I error does not exceed the 95% level.

Appendix E: Weak $A \to X$ and bias-variance tradeoff

First, we analyze the behavior of our methods for varying strength from A to X. For $N = 200$, we consider the coverage and length of the confidence intervals for varying strength from A to X for the same settings as in Figure 2 and 5.

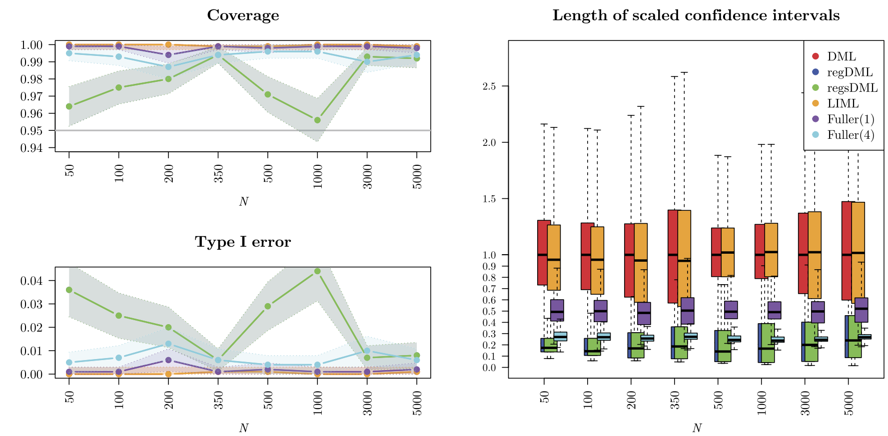

FIG 7. *The results come from $M = 1000$ simulation runs each from the SEM in Figure 1 with $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0: \ \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines as well as the red and orange ones are indistinguishable in the left panel.*

Figure 9 illustrates the results for data from the SEM from Figure 2. We vary the strength of the direct link $A \to X$ and denote it by α in Figure 9. Figure 10 illustrates the results for data from the SEM from Figure 5. We leave the link $A_2 \to X$ as it is and only vary the strength of the direct link $A_1 \to X$, which we denote by α in Figure 10. In both Figure 9 and 10, the coverage remains high for all considered methods. If α becomes larger in absolute value, the confidence intervals become shorter, which leads to a coverage that is closer to the nominal 95% level, especially in Figure 10. The regsDML method yields the shortest confidence intervals in both figures.

Second, we analyze the bias-variance tradeoff of the respective estimated quantities of the regularized methods. We again choose the sample size $N = 200$ and consider the same settings as in Figure 2 and 5. The results are summarized in Figure 11 and 12 that display the estimated MSE, estimated variance, and estimated squared bias as used in Equation (16). The MSE in both figures is mainly driven by the variance, and regsDML achieves a considerable variance reduction compared to the TSLS-type DML estimator.

Appendix F: Confounding and its mitigation

If we say in this section that the nuisance parameters are estimated with additive splines, they are estimated with additive cubic B-splines with $\lceil N^{\frac{1}{5}} \rceil + 2$

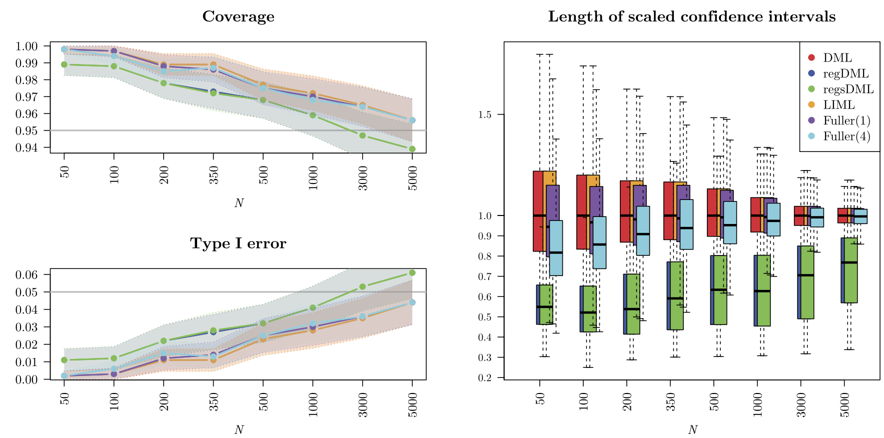

FIG 8. *The results come from* $M = 1000$ *simulation runs from the SEM in Figure 4 with* $\beta_0 = 0$ *for a range of sample sizes* N *and with* $K = 2$ *and* $\mathcal{S} = 100$ *in Algorithm 1. The nuisance functions are estimated with random forests. The figure displays the coverage of two-sided confidence intervals for* β_0*, type I error for two-sided testing of the hypothesis* $H_0: \ \beta_0 = 0$*, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level* 95%*. At each sample size* N*, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent* 95% *confidence bands with respect to the* M *simulation runs. The blue and green lines as well as the red and orange ones are indistinguishable in the left panel.*

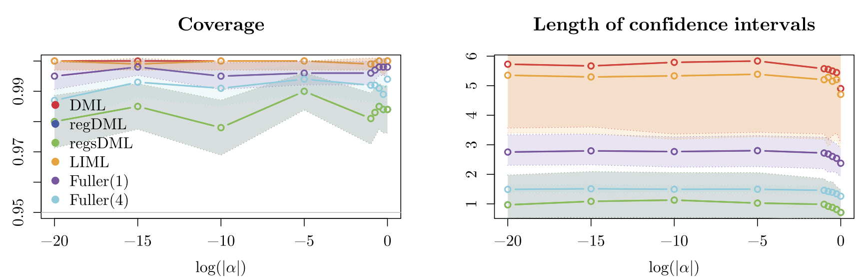

FIG 9. *Same setting as in Figure 2, but with* $N = 200$ *only. The strength of the direct link* $A \to X$ *varies and is denoted by* α*. We considered the* α*-values* $-e^{-20}$, $-e^{-15}$, $-e^{-10}$, $-e^{-5}$, $-e^{-1}$, $-e^{-0.75}$, $-e^{-0.5}$, $-e^{-0.25}$*, and* $-e^{0}$.

degrees of freedom, where N denotes the sample size of the data.

We consider models where the DML and the regsDML methods do not work well in terms of coverage of β_0. We present possible explanations of these failures and illustrate model changes to overcome them. The first model in Section F.1 features a strong confounding effect $H \to X$, the second model in Section F.2 features an effect with noise in $W \to H$, and the third model in Section F.3 features an effect with noise in $H \to W$.

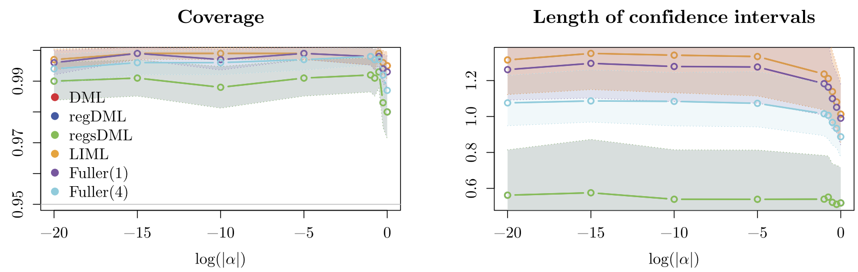

FIG 10. *Same setting as in Figure 5, but with $N = 200$ only. The strength of the direct link $A_1 \to X$ varies and is denoted by α. We considered the α-values e^{-20}, e^{-15}, e^{-10}, e^{-5}, e^{-1}, $e^{-0.75}$, $e^{-0.5}$, $e^{-0.25}$, and e^{0}.*

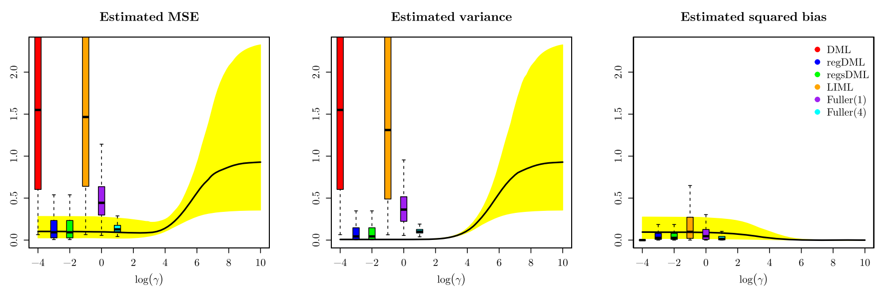

FIG 11. *Estimated MSE, estimated variance, and estimated squared bias as used in Equation (16) for the same setting as in Figure 2, but with $N = 200$ only. The black solid line displays the median of the respective quantity over the considered range of γ-values for $\hat{b}^{\gamma}$. The yellow area marks the observed 25% and 75% quantiles. All methods incorporate an additional variance adjustment from the $\mathcal{S}$ repetitions according to Algorithm 1. Boxplots illustrate the performance of the TSLS and the regularized methods. The position of the boxplots is not linked to the γ-values on the x-axis.*

F.1. Strong confounding effect $H \to X$

If the hidden variable H is strongly confounded with X, the resulting TSLS-type DML estimator can be substantially biased depending on the choice of functions in the model. If the estimated variances are not large enough, the coverage of the resulting confidence intervals for β_0 can be too low. This issue is illustrated in Figure 14.

The regsDML estimator mimics the bias behavior of DML because the DML estimator is used as a replacement of β_0 in the MSE objective function that defines the estimated regularization parameter of regDML in (16). The confidence intervals of regsDML are shorter than the DML ones, but both are computed with a similarly biased coefficient estimate of β_0. Therefore, the coverage of the confidence intervals of regsDML is even worse than the one of DML.

The coverages of both DML and regsDML are considerably improved if the confounding strength is reduced; see Figure 15.

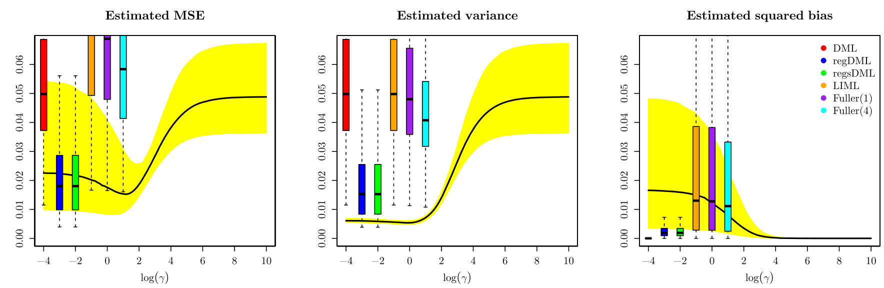

Fig 12. *Estimated MSE, estimated variance, and estimated squared bias as used in Equation* (16) *for the same setting as in Figure 5, but with* $N = 200$ *only. The black solid line displays the median of the respective quantity over the considered range of* γ*-values for* $\hat{b}^{\gamma}$. *The yellow area marks the observed* 25% *and* 75% *quantiles. All methods incorporate an additional variance adjustment from the* $\mathcal{S}$ *repetitions according to Algorithm 1. Boxplots illustrate the performance of the TSLS and the regularized methods. The position of the boxplots is not linked to the* γ*-values on the* x*-axis.*

$$
\begin{aligned}
&(\varepsilon_A, \varepsilon_W, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_5(\mathbf{0}, \mathbb{1})\\
&A \ \leftarrow \varepsilon_A\\
&W \leftarrow \varepsilon_W\\
&H \ \leftarrow \varepsilon_H\\
&X \ \leftarrow A + W + \chi H + 0.25\varepsilon_X\\
&Y \ \leftarrow \beta_0 X + W + H + 0.25\varepsilon_Y
\end{aligned}
$$

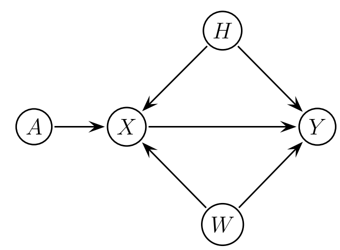

Fig 13. *An SEM and its associated causal graph.*

F.2. Noise in $W \to H$

The variable W may have a direct effect on H. If this link is strong enough with respect to the additional noise ε_H of H, it is possible to obtain some information of H by observing W. This can reduce the overall level of confounding present depending on the choice of functions in the model.

Simulation results where W explains only part of the variation in H are presented in Figure 17. The confidence intervals of both DML and regsDML do not attain a 95% coverage for small sample sizes N. The situation can be considerably improved by reducing the variation of H that is not explained by W; see Figure 18.

F.3. Noise in $H \to W$

The variable H may have a direct effect on W. If this link is strong enough with respect to the additional noise ε_W of W, it is possible to obtain some information of H by observing W similarly to Section F.2. The results again

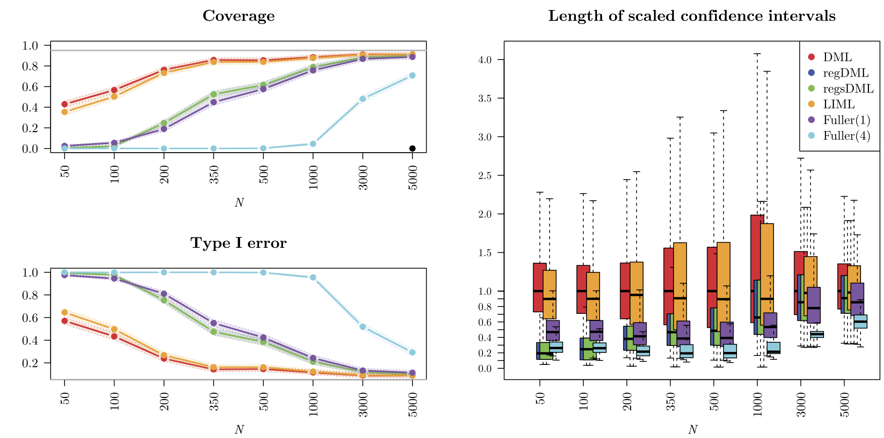

FIG 14. *The results come from $M = 1000$ simulation runs from the SEM in Figure 13 with $\chi = 15$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0 : \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

depend on the choice of functions in the model.

Figure 20 presents simulation results where H explains only little variation of W compared with ε_W. The confidence intervals of regsDML do not attain a 95% coverage for small sample sizes N because the estimator inherits additional bias from DML. The situation can be improved by reducing the variation of W that is not explained by H; see Figure 21.

Appendix G: Examples where the identifiability condition (5) does and does not hold

The following examples illustrate SEMs where the identifiability condition (5) holds and where it fails to hold. We argue using causal graphs; see Lauritzen [59], Pearl [74, 76, 77], Peters, Janzing and Schölkopf [78], and Maathuis et al. [64]. By convention, we omit error variables in a causal graph if they are assumed to be mutually independent [76].

Example G.1. *Consider the SEM of the 1-dimensional variables A, W, H, X, and Y and its associated causal graph given in Figure 22, where β_0 is a fixed unknown parameter, and where a_W, a_X, g_Y, g_H, h_X, and h_Y are some appropriate functions. The variable A directly influences W, and W directly influences the hidden variable H. The variable A is independent of H given W because every path from A to H is blocked by W.*

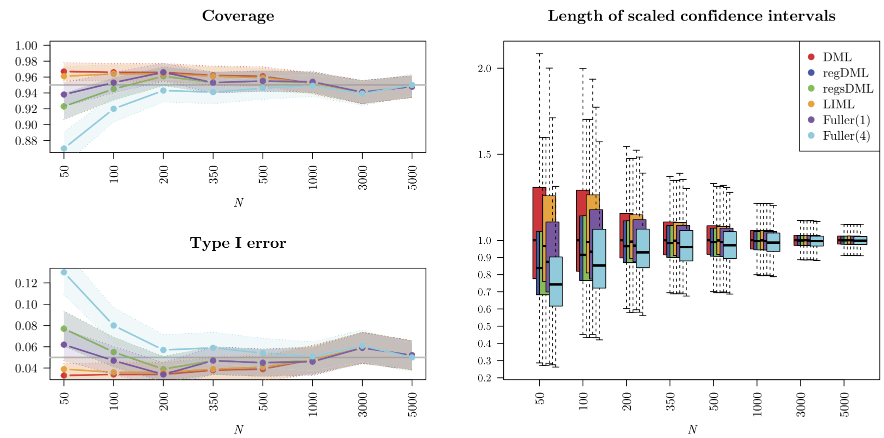

FIG 15. *The results come from $M = 1000$ simulation runs from the SEM in Figure 13 with $\chi = 1$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0: \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

$$
\begin{aligned}
&(\varepsilon_A, \varepsilon_W, \varepsilon_H, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_5(\mathbf{0}, \mathbb{1}) \\
&A \leftarrow \varepsilon_A \\
&W \leftarrow \varepsilon_W \\
&H \leftarrow W + \kappa\varepsilon_H \\
&X \leftarrow 0.5A + 3\tanh(2W) + 1.5H + 0.25\varepsilon_X \\
&Y \leftarrow \beta_0 X - \tanh(W) + H + 0.25\varepsilon_Y
\end{aligned}
$$

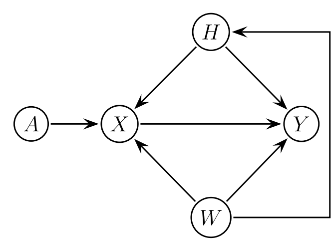

FIG 16. *An SEM and its associated causal graph.*

Proof of Example G.1. The path $A \to X \leftarrow H$ is blocked by the empty set because X is a collider on this path. The paths $A \to \cdots \to Y \leftarrow H$ are blocked by the empty set because Y is a collider on these paths. The path $A \to W \to H$ is blocked by W. □

The variable A is exogenous in Example G.1. In general, this is no requirement; see Example G.2.

Example G.2. *Consider the SEM of the 1-dimensional variables H, W, A, X, and Y and its associated causal graph given in Figure 23, where β_0 is a fixed unknown parameter, and where a_X, g_A, g_X, g_Y, h_X, h_W, and h_Y are some appropriate functions. The variable A is not a source node. The hidden*

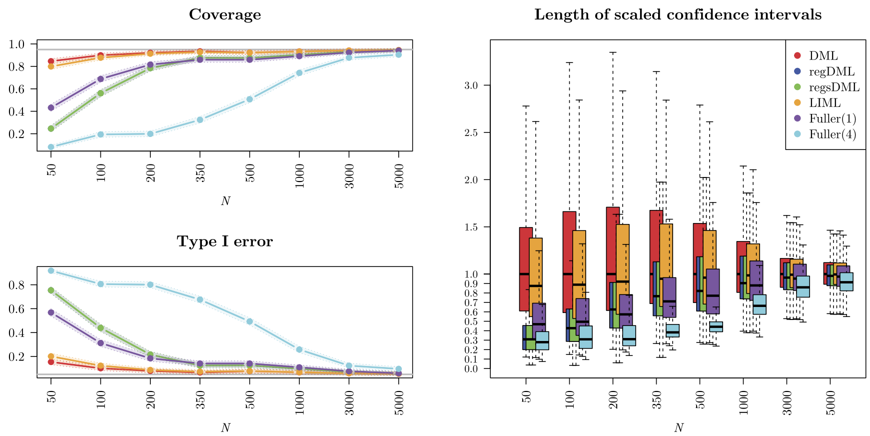

FIG 17. *The results come from $M = 1000$ simulation runs from the SEM in Figure 16 with $\kappa = 2$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0:\ \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

variable H directly influences W, and W directly influences A. The variable A is independent of H given W because every path from A to H is blocked by W.

Proof of Example G.2. The path $A \to X \leftarrow H$ is blocked by the empty set because X is a collider on this path. The paths $A \to X \to \cdots \to Y \leftarrow H$ are blocked by the empty set because Y is a collider on these paths. The paths $A \leftarrow W \to Y \leftarrow X \leftarrow H$, $A \leftarrow W \leftarrow H$, and $A \to X \leftarrow W \leftarrow H$ are blocked by W. The path $A \leftarrow W \to Y \leftarrow H$ is blocked by W or alternatively by the empty set because Y is a collider on this path. The path $A \leftarrow W \to X \leftarrow H$ is blocked by W or alternatively by the empty set because X is a collider on this path. □

Identifiability of β_0 is not guaranteed if A and H are independent. An illustration is given in Example G.3. Considering the instrument A instead of R_A in Theorem 2.1 cannot solve the issue. In such a situation, stronger structural assumptions are required.

Example G.3. *Consider the SEM of the 1-dimensional variables H, A, W, X, and Y and its associated causal graph given in Figure 24, where β_0 is a fixed unknown parameter. Although A and H are independent, the identifiability condition (5) does not hold.*

Proof of Example G.3. The two random variables A and H are independent

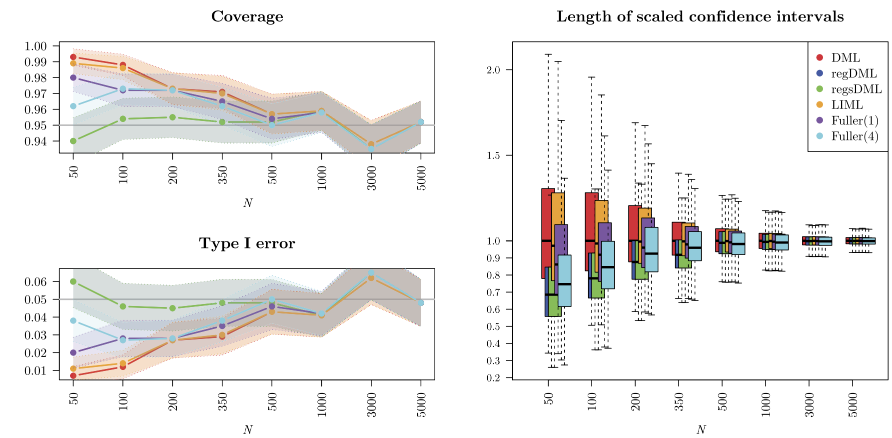

FIG 18. *The results come from $M = 1000$ simulation runs from the SEM in Figure 16 with $\kappa = 0.25$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0: \ \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%, and where the nuisance functions are estimated with additive splines. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

$$
\begin{aligned}
&(\varepsilon_H, \varepsilon_W, \varepsilon_A, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_5(\mathbf{0}, \mathbb{1}) \\
&H \leftarrow \varepsilon_H \\
&W \leftarrow 2H + \kappa\varepsilon_W \\
&A \leftarrow e^{-0.5W} + 0.5\varepsilon_A \\
&X \leftarrow -A - 0.1W^3 - 0.2W^2 + 0.4W \\
&\qquad\quad + \frac{7}{1+e^{-4H}} + 0.25\varepsilon_X \\
&Y \leftarrow \beta_0 X + 0.5W + 0.5H + \varepsilon_Y
\end{aligned}
$$

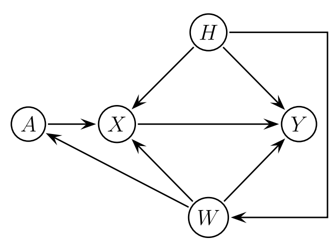

FIG 19. *An SEM and its associated causal graph.*

because the path $A \to W \leftarrow H$ is not blocked by W. Indeed, W is a collider on this path.

All random variables are 1-dimensional. Therefore, the representation of β_0 in Theorem 2.1 is equivalent to the identifiability condition

$$\mathbb{E}[R_A(R_Y - R_X\beta_0)] = 0$$

in Equation (5). However, the identifiability condition does not hold in the

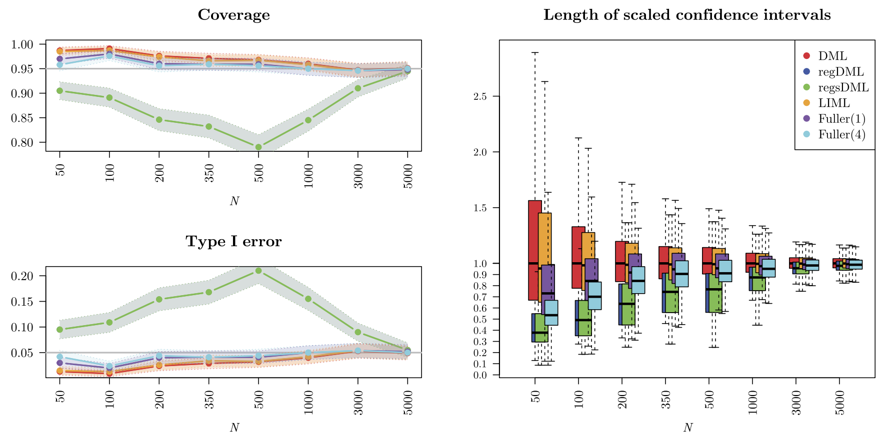

FIG 20. *The results come from $M = 1000$ simulation runs from the SEM in Figure 19 with $\kappa = 1$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0 : \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

present situation. We have

$$\begin{aligned}&\mathbb{E}[R_A(R_Y - R_X\beta_0)]\\ &= \mathbb{E}[R_A\big(H + \varepsilon_Y - \mathbb{E}[H + \varepsilon_Y|W]\big)]\\ &= \mathbb{E}\left[R_A\big(H - \mathbb{E}[H|W]\big)\right]\end{aligned}$$

because ε_Y is independent of A and W and centered. By the tower property for conditional expectations, we have

$$\mathbb{E}[R_A(R_Y - R_X\beta_0)] = \mathbb{E}\left[AH - A\,\mathbb{E}[H|W]\right].$$

Because A and H are independent and centered, we have $\mathbb{E}[AH] = 0$. Moreover, we have $H \sim \mathcal{N}(0,1)$, $W \sim \mathcal{N}(0,3)$, and $(W|H = h) \sim \mathcal{N}(h,2)$. The conditional distribution of $H|W = w$ can be obtained by applying Bayes' theorem and is given by $\mathcal{N}(\frac{1}{3}w, \frac{2}{3})$. Hence, we have $\mathbb{E}[H|W] = \frac{1}{3}W$ and

$$\mathbb{E}\left[A\,\mathbb{E}[H|W]\right] = \frac{1}{3}\,\mathbb{E}[AW] = \frac{1}{3}\,\mathbb{E}\left[A^2\right] = \frac{1}{3} \neq 0$$

because A is independent of H and ε_W. Therefore, we have $\mathbb{E}[R_A(R_Y - R_X\beta_0)] \neq 0$, and β_0 cannot be represented as in Theorem 2.1. □

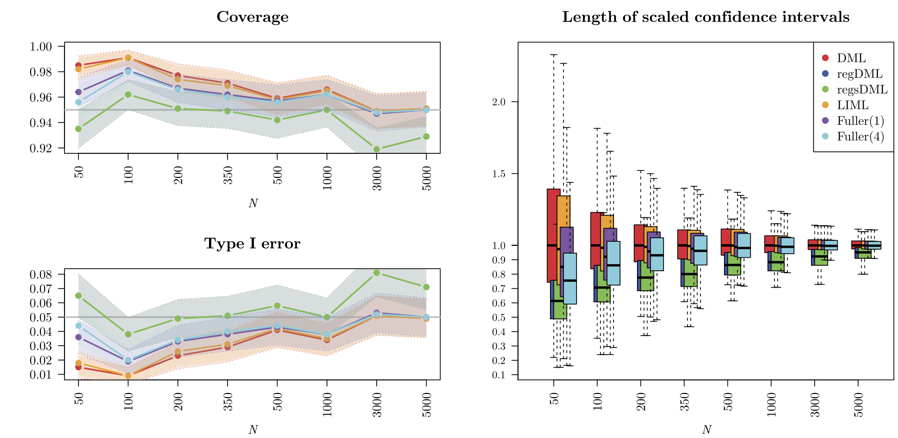

FIG 21. *The results come from $M = 1000$ simulation runs from the SEM in Figure 19 with $\kappa = 0.25$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 1. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence intervals for β_0, type I error for two-sided testing of the hypothesis $H_0: \ \beta_0 = 0$, and scaled lengths of two-sided confidence intervals of DML (red), regDML (blue), regsDML (green), LIML (orange), Fuller(1) (purple), and Fuller(4) (cyan), where all results are at level 95%. At each sample size N, the lengths of the confidence intervals are scaled with the median length from DML. The shaded regions in the coverage and the type I error plots represent 95% confidence bands with respect to the M simulation runs. The blue and green lines are indistinguishable in the left panel.*

$$
\begin{aligned}
&\varepsilon_A, \varepsilon_W, \varepsilon_H, \varepsilon_X, \varepsilon_Y \\
&A \leftarrow \varepsilon_A \\
&W \leftarrow a_W(A) + \varepsilon_W \\
&H \leftarrow g_H(W) + \varepsilon_H \\
&X \leftarrow a_X(A) + h_X(H) + \varepsilon_X \\
&Y \leftarrow \beta_0 X + g_Y(W) + h_Y(H) + \varepsilon_Y
\end{aligned}
$$

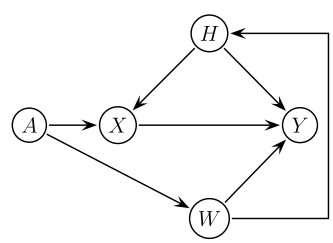

FIG 22. *An SEM satisfying the identifiability condition (5) and its associated causal graph as in Example G.1.*

Appendix H: Proofs of Section 2

Proof of Theorem 2.1. To prove the theorem, we need to verify that the representation

$$\beta_0 = \left(\mathbb{E}\left[R_X R_A^T\right] \mathbb{E}\left[R_A R_A^T\right]^{-1} \mathbb{E}\left[R_A R_X^T\right]\right)^{-1} \mathbb{E}\left[R_X R_A^T\right] \mathbb{E}\left[R_A R_A^T\right]^{-1} \mathbb{E}[R_A R_Y]$$

holds. This statement is equivalent to

$$\mathbf{0} = \mathbb{E}\left[R_X R_A^T\right] \mathbb{E}\left[R_A R_A^T\right]^{-1} \mathbb{E}\left[R_A\left(R_Y - R_X^T \beta_0\right)\right].$$

$$
\begin{aligned}
&\varepsilon_H, \varepsilon_W, \varepsilon_A, \varepsilon_X, \varepsilon_Y \\
&H \leftarrow \varepsilon_H \\
&W \leftarrow h_W(H) + \varepsilon_W \\
&A \leftarrow g_A(W) + \varepsilon_A \\
&X \leftarrow a_X(A) + g_X(W) + h_X(H) + \varepsilon_X \\
&Y \leftarrow \beta_0 X + g_Y(W) + h_Y(H) + \varepsilon_Y
\end{aligned}
$$

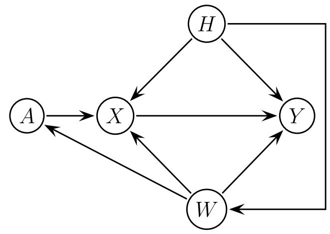

FIG 23. *An SEM satisfying the identifiability condition* (5) *and its associated causal graph as in Example G.2.*

$$
\begin{aligned}
&(\varepsilon_H, \varepsilon_A, \varepsilon_W, \varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_5(0, \mathbb{1}) \\
&H \leftarrow \varepsilon_H \\
&A \leftarrow \varepsilon_A \\
&W \leftarrow A + H + \varepsilon_W \\
&X \leftarrow A + W + H + \varepsilon_X \\
&Y \leftarrow \beta_0 X + W + H + \varepsilon_Y
\end{aligned}
$$

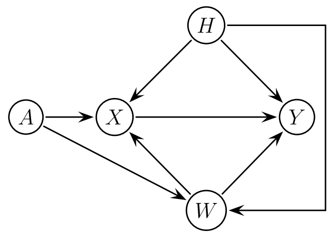

FIG 24. *An SEM not satisfying the identifiability condition* (5) *together with its associated causal graph as in Example G.3*

This last statement holds because $\mathbb{E}[R_A(R_Y - R_X^T\beta_0)]$ equals $\mathbf{0}$ due to the identifiability condition (5). ☐

Appendix I: Proofs of Section 3

We denote by $\|\cdot\|$ either the Euclidean norm for a vector or the operator norm for a matrix.

Proof of Proposition 3.3. We have

$$
\begin{aligned}
&\frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\left[\psi\big(S; \beta_0, \eta^0 + r(\eta - \eta^0)\big)\right] \\
&= \frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\Bigg[\Big(A - m_A^0(W) - r\big(m_A(W) - m_A^0(W)\big)\Big) \\
&\qquad\qquad \cdot\Bigg(Y - m_Y^0(W) - r\big(m_Y(W) - m_Y^0(W)\big) \\
&\qquad\qquad\qquad - \Big(X - m_X^0(W) - r\big(m_X(W) - m_X^0(W)\big)\Big)^T \beta_0\Bigg)\Bigg] \\
&= \mathbb{E}_P\Big[-\big(m_A(W) - m_A^0(W)\big)\Big(Y - m_Y^0(W) - \big(X - m_X^0(W)\big)^T\beta_0\Big) \\
&\qquad + \big(A - m_A^0(W)\big)\Big(-\big(m_Y(W) - m_Y^0(W)\big) \\
&\qquad\qquad\qquad + \big(m_X(W) - m_X^0(W)\big)^T\beta_0\Big)\Big].
\end{aligned}
$$

Subsequently, we show that both terms

$$\mathbb{E}_P\left[\left(m_A(W)-m_A^0(W)\right)\left(Y-m_Y^0(W)-\left(X-m_X^0(W)\right)^T\beta_0\right)\right] \tag{21}$$

and

$$\mathbb{E}_P\left[\left(A-m_A^0(W)\right)\left(-\left(m_Y(W)-m_Y^0(W)\right)+\left(m_X(W)-m_X^0(W)\right)^T\beta_0\right)\right] \tag{22}$$

are equal to $\mathbf{0}$. We first consider the term (21). Recall the notations $m_Y^0(W)=\mathbb{E}_P[Y|W]$ and $m_X^0(W)=\mathbb{E}_P[X|W]$. We have

$$\begin{aligned}
&\mathbb{E}_P\left[\left(m_A(W)-m_A^0(W)\right)\left(Y-m_Y^0(W)-\left(X-m_X^0(W)\right)^T\beta_0\right)\right]\\
&=\mathbb{E}_P\left[\left(m_A(W)-m_A^0(W)\right)\mathbb{E}_P\left[Y-\mathbb{E}_P[Y|W]-(X-\mathbb{E}_P[X|W])^T\beta_0\middle|W\right]\right]\\
&=\mathbf{0}.
\end{aligned}$$

Next, we verify that the term given in (22) vanishes. Recall that we denote $m_A^0(W)=\mathbb{E}_P[A|W]$. We have

$$\begin{aligned}
&\mathbb{E}_P\left[\left(A-m_A^0(W)\right)\left(-\left(m_Y(W)-m_Y^0(W)\right)+\left(m_X(W)-m_X^0(W)\right)^T\beta_0\right)\right]\\
&=\mathbb{E}_P\Big[\mathbb{E}_P\left[A-\mathbb{E}[A|W]\middle|W\right]\Big(-\left(m_Y(W)-m_Y^0(W)\right)\\
&\qquad\qquad+\left(m_X(W)-m_X^0(W)\right)^T\beta_0\Big)\Big]\\
&=\mathbf{0}.
\end{aligned}$$

Because both terms (21) and (22) vanish, we conclude

$$\frac{\partial}{\partial r}\Big|_{r=0}\mathbb{E}_P\left[\psi\big(S;\beta_0,\eta^0+r(\eta-\eta^0)\big)\right]=\mathbf{0}. \qquad \square$$

Definition I.1. *Consider a set $\mathcal{T}$ of nuisance functions. For $S=(A,X,W,Y)$, an element $\eta=(m_A,m_X,m_Y)\in\mathcal{T}$, and $\beta\in\mathbb{R}^d$, we introduce the score functions*

$$\widetilde{\psi}(S,\beta,\eta):=\left(X-m_X(W)\right)\left(Y-m_Y(W)-\left(X-m_X(W)\right)^T\beta\right), \tag{23}$$

and

$$\begin{aligned}
\psi_1(S,\eta)&:=\left(X-m_X(W)\right)\left(A-m_A(W)\right)^T,\\
\psi_2(S,\eta)&:=\left(A-m_A(W)\right)\left(A-m_A(W)\right)^T,\\
\psi_3(S,\eta)&:=\left(X-m_X(W)\right)\left(X-m_X(W)\right)^T.
\end{aligned}$$

Furthermore, let the matrices

$$\begin{aligned}
D_1&:=\mathbb{E}_P[\psi_3(S;\eta^0)],\\
D_2&:=\mathbb{E}_P[\psi_1(S;\eta^0)]\,\mathbb{E}_P[\psi_2(S;\eta^0)]^{-1}\,\mathbb{E}_P\left[\psi_1^T(S;\eta^0)\right],\\
D_3&:=\mathbb{E}_P[\psi_1(S;\eta^0)]\,\mathbb{E}_P[\psi_2(S;\eta^0)]^{-1},\\
D_5&:=\mathbb{E}_P[\psi_2(S;\eta^0)]^{-1}\,\mathbb{E}_P[\psi(S;b^\gamma,\eta^0)],\\
J_0&:=D_2^{-1}D_3,\\
\tilde{J}_0&:=\mathbb{E}_P\left[\psi(S;\beta_0,\eta^0)\psi^T(S;\beta_0,\eta^0)\right]=\mathbb{E}_P\left[R_AR_A^T(R_Y-R_X^T\beta_0)^2\right],\\
J_0''&:=\mathbb{E}_P[R_AR_A^T],\\
J_0'&:=\mathbb{E}_P\left[R_X(R_A)^T\right](J_0'')^{-1}\,\mathbb{E}_P\left[R_A(R_X)^T\right]
\end{aligned}$$

and the variance-covariance matrix $\sigma^2 := J_0 \tilde{J}_0 J_0^T$. *Moreover, let the score function*

$$\overline{\psi}(\cdot; \beta_0, \eta^0) := \sigma^{-1} \tilde{J}_0^{-\frac{1}{2}} \psi(\cdot; \beta_0, \eta^0).$$

Definition I.2. *Let* $\gamma \geq 0$. *Consider a realization set* $\mathcal{T}$ *of nuisance functions. Define the statistical rates*

$$r_N^4 := \max_{\substack{S=(U,V,W,Z) \in \{A,X,Y\}^2 \times \{W\} \times \{A,X,Y\}, \\ b^0 \in \{b^\gamma, \beta_0, \mathbf{0}\}}} \sup_{\eta \in \mathcal{T}} \mathbb{E}_P[\|\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0)\|]$$

and

$$\lambda_N := \max_{\substack{\varphi \in \{\psi, \widetilde{\psi}, \psi_2\}, \\ b^0 \in \{b^\gamma, \beta_0, \mathbf{0}\}}} \sup_{r \in (0,1), \eta \in \mathcal{T}} \left\| \partial_r^2 \, \mathbb{E}_P \left[\varphi\big(S; b^0, \eta^0 + r(\eta - \eta^0)\big) \right] \right\|,$$

where we interpret $\psi_2\big(S; b^0, \eta^0 + r(\eta - \eta^0)\big)$ *as* $\psi_2\big(S; \eta^0 + r(\eta - \eta^0)\big)$ *in the definition of* λ_N.

Remark I.3. *We would like to remark that the respective definition of the statistical rate* r_N *given in Chernozhukov et al. [31] involves the* L_2*-norm of* $\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0)$ *instead of its* L_1*-norm. However, it is essential to employ the* L_1*-norm to ensure that Assumption I.5.5 can constrain the* L_2*-norm of the estimation errors incurred by the ML estimators of the nuisance parameters. Thus, we do not have to constrain their higher order errors to employ Hölder's inequality in Lemma I.16.*

Definition I.4. *Let the nonrandom numbers*

$$\rho_N := r_N + N^{\frac{1}{2}} \lambda_N \quad \text{and} \quad \tilde{\rho}_N := N^{\max\left\{\frac{4}{p} - 1, -\frac{1}{2}\right\}} + r_N.$$

If not stated otherwise, we assume the following Assumption I.5 in all the results presented in the appendix.

Assumptions I.5. *Let* $\gamma \geq 0$. *Let* $K \geq 2$ *be a fixed integer independent of* N. *We assume that* $N \geq K$ *holds. Let* $\{\delta_N\}_{N \geq K}$ *and* $\{\Delta_N\}_{N \geq K}$ *be two sequences of positive numbers that converge to zero, where* $\delta_N^{\frac{1}{4}} \geq N^{-\frac{1}{2}}$ *holds. Let* $\{\mathcal{P}_N\}_{N \geq 1}$ *be a sequence of sets of probability distributions* P *of the quadruple* $S = (A, W, X, Y)$.

Let $p > 4$. *For all* N, *for all* $P \in \mathcal{P}_N$, *consider a nuisance function realization set* $\mathcal{T}$ *such that the following conditions hold:*

- *I.5.1 We have an SEM given by (3) that satisfies the identifiability conditon (5).*
- *I.5.2 There exists a finite real constant* C_1 *satisfying* $\|A\|_{P,p} + \|X\|_{P,p} + \|Y\|_{P,p} \leq C_1$.
- *I.5.3 The matrix* $\mathbb{E}_P[R_X R_A^T] \in \mathbb{R}^{d \times q}$ *has full rank* d. *This in particular requires* $q \geq d$. *The matrices* $D_1 \in \mathbb{R}^{d \times d}$ *and* $J_0'' \in \mathbb{R}^{q \times q}$ *are invertible. Furthermore, the smallest and largest singular values of the symmetric matrices* J_0'' *and* J_0' *are bounded away from* 0 *by* $c_1 > 0$ *and are bounded away from* $+\infty$ *by* $c_2 < \infty$.

I.5.4 *The symmetric matrices* $\tilde{J}_0$, $D_1 + (\gamma - 1)D_2$, *and* D_4 *are invertible, where* D_4 *is introduced in Definition J.1 in the appendix in Section J. The smallest and largest singular values of these matrices are bounded away from* 0 *by* c_3 *and are bounded away from* $+\infty$ *by* c_4.

I.5.5 *The set* $\mathcal{T}$ *consists of* P*-integrable functions* $\eta = (m_A, m_X, m_Y)$ *whose pth moment exists and it contains* η^0. *There exists a finite real constant* C_2 *such that*

$$\begin{aligned}
&\|\eta^0 - \eta\|_{P,p} \le C_2, \quad \|\eta^0 - \eta\|_{P,2} \le \delta_N, \\
&\|m_A^0(W) - m_A(W)\|_{P,2}^2 \le \delta_N N^{-\frac{1}{2}}, \\
&\|m_X^0(W) - m_X(W)\|_{P,2} \\
&\quad \cdot \big(\|m_Y^0(W) - m_Y(W)\|_{P,2} + \|m_X^0(W) - m_X(W)\|_{P,2}\big) \le \delta_N N^{-\frac{1}{2}}, \\
&\|m_A^0(W) - m_A(W)\|_{P,2} \\
&\quad \cdot \big(\|m_Y^0(W) - m_Y(W)\|_{P,2} + \|m_X^0(W) - m_X(W)\|_{P,2}\big) \le \delta_N N^{-\frac{1}{2}}
\end{aligned}$$

hold for all elements η *of* $\mathcal{T}$. *Given a partition* $I_1, \ldots, I_K$ *of* $[N]$ *where each* I_k *is of size* $n = \frac{N}{K}$, *for all* $k \in [K]$, *the nuisance parameter estimate* $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\{S_i\}_{i \in I_k^c})$ *satisfies*

$$\begin{aligned}
&\|\eta^0 - \hat{\eta}^{I_k^c}\|_{P,p} \le C_2, \quad \|\eta^0 - \hat{\eta}^{I_k^c}\|_{P,2} \le \delta_N, \\
&\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2}^2 \le \delta_N N^{-\frac{1}{2}}, \\
&\|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2} \\
&\quad \cdot \big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \le \delta_N N^{-\frac{1}{2}}, \\
&\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2} \\
&\quad \cdot \big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \le \delta_N N^{-\frac{1}{2}}
\end{aligned}$$

with P*-probability no less than* $1 - \Delta_N$. *Denote by* $\mathcal{E}_N$ *the event that* $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\{S_i\}_{i \in I_k^c})$ *belongs to* $\mathcal{T}$, *and assume that this event holds with* P*-probability no less than* $1 - \Delta_N$.

For instance, invertibility of the square matrices $\mathbb{E}_P[R_A R_A^T]$ and $\tilde{J}_0$ is satisfied if ε_Y is independent of both A and W and has a strictly positive variance.

Remark I.6. *It is possible to drop some of the assumptions in Assumption I.5 if we are interested in proving the results about DML only. The full assumption is required to prove the results about DML and the regularized methods.*

Lemma I.7. *Let* $u \ge 1$. *Consider a* t*-dimensional random variable* Z. *Denote the joint law of* Z *and* W *by* P. *Then we have*

$$\|Z - \mathbb{E}_P[Z|W]\|_{P,u} \le 2\|Z\|_{P,u}.$$

Proof of Lemma I.7. Because the Euclidean norm to the uth power is convex for $u \ge 1$, we have

$$\begin{aligned}
&\|\mathbb{E}_P[Z|W]\|_{P,u}^u \\
&= \mathbb{E}_P\left[\|\mathbb{E}_P[Z|W]\|^u\right] \\
&\le \mathbb{E}_P\left[\mathbb{E}_P[\|Z\|^u|W]\right] \\
&= \mathbb{E}_P[\|Z\|^u] \\
&= \|Z\|_{P,u}^u
\end{aligned}$$

by Jensen's inequality. We hence have

$$\|Z - \mathbb{E}_P[Z|W]\|_{P,u} \leq \|Z\|_{P,u} + \|\mathbb{E}_P[Z|W]\|_{P,u} \leq 2\|Z\|_{P,u}$$

by the triangle inequality. □

Lemma I.8. *Consider a t-dimensional random variable Z. Denote the joint law of Z and W by P. Then we have*

$$\left\| \mathbb{E}_P \left[ZZ^T - \mathbb{E}_P[Z|W]\,\mathbb{E}_P[Z^T|W]\right] \right\| \leq 2\|Z\|_{P,2}^2.$$

Proof of Lemma I.8. Because the Euclidean norm is convex, we have

$$\begin{aligned} &\left\| \mathbb{E}_P \left[ZZ^T - \mathbb{E}_P[Z|W]\,\mathbb{E}_P[Z^T|W]\right] \right\| \\ &\leq \mathbb{E}_P \left[\|ZZ^T\| + \|\mathbb{E}_P[Z|W]\,\mathbb{E}_P[Z^T|W]\|\right] \\ &\leq \mathbb{E}_P \left[\|Z\|^2 + \|\mathbb{E}_P[Z|W]\|^2\right] \end{aligned}$$

by Jensen's inequality, the triangle inequality, and the Cauchy–Schwarz inequality. Because the squared Euclidean norm is convex, we have

$$\|\mathbb{E}_P[Z|W]\|^2 \leq \mathbb{E}_P \left[\|Z\|^2 \middle| W\right]$$

by Jensen's inequality. Therefore, we have

$$\begin{aligned} &\left\| \mathbb{E}_P \left[ZZ^T - \mathbb{E}_P[Z|W]\,\mathbb{E}_P[Z^T|W]\right] \right\| \\ &\leq \mathbb{E}_P \left[\|Z\|^2 + \|\mathbb{E}_P[Z|W]\|^2\right] \\ &\leq \mathbb{E}_P \left[\|Z\|^2 + \mathbb{E}_P[\|Z\|^2|W]\right] \\ &= 2\|Z\|_{P,2}^2. \quad \square \end{aligned}$$

Lemma I.9. *Let a t_1-dimensional random variable Z_1 and a t_2-dimensional random variable Z_2. Denote the joint law of Z_1, Z_2, and W by P. Then we have*

$$\left\| \mathbb{E}_P \left[(Z_1 - \mathbb{E}_P[Z_1|W])(Z_2 - \mathbb{E}_P[Z_2|W])^T\right] \right\|^2 \leq \|Z_1\|_{P,2}^2 \|Z_2\|_{P,2}^2.$$

Proof of Lemma I.9. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} &\left\| \mathbb{E}_P \left[(Z_1 - \mathbb{E}_P[Z_1|W])(Z_2 - \mathbb{E}_P[Z_2|W])^T\right] \right\|^2 \\ &\leq \mathbb{E}_P \left[\|(Z_1 - \mathbb{E}_P[Z_1|W])\|^2\right] \mathbb{E}_P \left[\|(Z_2 - \mathbb{E}_P[Z_2|W])\|^2\right]. \end{aligned}$$

Because the conditional expectation minimizes the mean squared error [39, Theorem 5.1.8], we have

$$\mathbb{E}_P \left[\|(Z_1 - \mathbb{E}_P[Z_1|W])\|^2\right] \leq \|Z_1\|_{P,2}^2$$

and

$$\mathbb{E}_P \left[\|(Z_2 - \mathbb{E}_P[Z_2|W])\|^2\right] \leq \|Z_2\|_{P,2}^2.$$

In total, we thus have

$$\left\| \mathbb{E}_P \left[(Z_1 - \mathbb{E}_P[Z_1|W])(Z_2 - \mathbb{E}_P[Z_2|W])^T\right] \right\|^2 \leq \|Z_1\|_{P,2}^2 \|Z_2\|_{P,2}^2. \qquad \square$$

Lemma I.10. *Let a t_1-dimensional random variable Z_1 and a t_2-dimensional random variable Z_2. Denote the joint law of Z_1, Z_2, and W by P. Then we have*

$$\big\| \mathbb{E}_P\left[(Z_1 - \mathbb{E}_P[Z_1|W])Z_2^T\right]\big\|^2 \le \|Z_1\|_{P,2}^2 \|Z_2\|_{P,2}^2.$$

Proof of Lemma I.10. By the Cauchy–Schwarz inequality, we have

$$\big\| \mathbb{E}_P\left[(Z_1 - \mathbb{E}_P[Z_1|W])Z_2^T\right]\big\|^2 \le \mathbb{E}_P\left[\|Z_1 - \mathbb{E}_P[Z_1|W]\|^2\right] \mathbb{E}_P\left[\|Z_2\|^2\right].$$

Because the conditional expectation minimizes the mean squared error [39, Theorem 5.1.8], we have

$$\mathbb{E}_P\left[\|Z_1 - \mathbb{E}_P[Z_1|W]\|^2\right] \le \mathbb{E}_P\left[\|Z_1\|^2\right] = \|Z_1\|_{P,2}^2.$$

Consequently,

$$\big\| \mathbb{E}_P\left[(Z_1 - \mathbb{E}_P[Z_1|W])Z_2^T\right]\big\|^2 \le \|Z_1\|_{P,2}^2 \|Z_2\|_{P,2}^2$$

holds. □

Lemma I.11. *Let $a, b \in \mathbb{R}$ be two numbers. We have*

$$(a+b)^2 \le 2a^2 + 2b^2. \tag{24}$$

Proof of Lemma I.11. The true statement $0 \le (a-b)^2$ is equivalent to (24). □

The following lemma proved in Chernozhukov et al. [31] states that conditional convergence in probability implies unconditional convergence in probability.

Lemma I.12 (Based on Chernozhukov et al. [31, Lemma 6.1].)**.** *Let $\{X_t\}_{t\ge1}$ and $\{Y_t\}_{t\ge1}$ be sequences of random vectors, and let $u \ge 1$. Consider a deterministic sequence $\{\varepsilon_t\}_{t\ge1}$ with $\varepsilon_t \to 0$ as $t \to \infty$ such that we have $\mathbb{E}[\|X_t\|^u|Y_t] \le \varepsilon_t^u$. Then we have $\|X_t\| = O_P(\varepsilon_t)$ unconditionally, meaning that that for any sequence $\{\ell_t\}_{t\ge1}$ with $\ell_t \to \infty$ as $t \to \infty$, we have $P(\|X_t\| > \ell_t\varepsilon_t) \to 0$.*

Proof of Lemma I.12. We have

$$P(\|X_t\| > \ell_t\varepsilon_t) = \mathbb{E}[P(\|X_t\| > \ell_t\varepsilon_t|Y_t)] \le \frac{\mathbb{E}\left[\mathbb{E}[\|X_t\|^u|Y_t]\right]}{\ell_t^u\varepsilon_t^u} \le \frac{1}{\ell_t^u} \to 0 \quad (t \to \infty)$$

by Markov's inequality. □

Lemma I.13. *There exists a finite real constant C_3 satisfying $\|\beta_0\| \le C_3$.*

Proof of Lemma I.13. Recall the matrices J_0' and J_0'' in Definition I.1. We have

$$\begin{array}{l} \|\beta_0\| \\ \le \big\|(J_0')^{-1}\big\|\big\|\mathbb{E}_P\left[AR_X^T\right]\big\|\big\|(J_0'')^{-1}\big\|\big\|\mathbb{E}_P\left[AR_Y\right]\big\| \\ \le \frac{1}{c_2^2}\|X\|_{P,2}\|Y\|_{P,2}\|A\|_{P,2}^2 \end{array}$$

by submultiplicativity, Assumption I.5.3, and Lemma I.10. We hence infer

$$\|\beta_0\| \leq \frac{1}{c_2^2} C_1^4$$

by Assumption I.5.2. □

Lemma I.14. *Let $\gamma \geq 0$. There exists a finite real constant C_4 satisfying $\|b^\gamma\| \leq C_4$.*

Proof of Lemma I.14. We have

$$\begin{aligned} &\|b^\gamma\| \\ &\leq \left\| \left(\mathbb{E}_P\left[R_X R_X^T\right] + (\gamma-1)\,\mathbb{E}_P\left[R_X R_A^T\right] \mathbb{E}_P\left[R_A R_A^T\right]^{-1} \mathbb{E}_P\left[R_A R_X^T\right] \right)^{-1} \right\| \\ &\quad \cdot \left\| \mathbb{E}_P[R_X R_Y] + (\gamma-1)\,\mathbb{E}_P\left[R_X R_A^T\right] \mathbb{E}_P\left[R_A R_A^T\right]^{-1} \mathbb{E}_P[R_A R_Y] \right\| \end{aligned}$$

by submultiplicativity. By Assumption I.5.4, the largest singular value of the matrix

$$D_1 + (\gamma-1)D_2 = \mathbb{E}_P\left[R_X R_X^T\right] + (\gamma-1)\,\mathbb{E}_P\left[R_X R_A^T\right] \mathbb{E}_P\left[R_A R_A^T\right]^{-1} \mathbb{E}_P\left[R_A R_X^T\right]$$

is upper bounded by $0 < c_4 < \infty$. Thus, we have

$$\begin{aligned} \|b^\gamma\| \leq \tfrac{1}{c_4}\Big(&\|\mathbb{E}_P[R_X R_Y]\| \\ &+ |\gamma-1| \left\| \mathbb{E}_P\left[R_X R_A^T\right] \right\| \left\| \mathbb{E}_P\left[R_A R_A^T\right]^{-1} \right\| \left\| \mathbb{E}_P\left[R_A R_Y^T\right] \right\| \Big) \end{aligned}$$

by the triangle inequality and submultiplicativity. By Assumption I.5.3, the largest singular value of $\mathbb{E}_P[R_A R_A^T]$ is upper bounded by $0 < c_2 < \infty$. By Lemma I.9 and Assumption I.5.2, we have

$$\begin{aligned} \left\| \mathbb{E}_P\left[R_X R_Y\right] \right\| &\leq \|X\|_{P,2} \|Y\|_{P,2} \leq C_1^2, \\ \left\| \mathbb{E}_P\left[R_X R_A^T\right] \right\| &\leq \|X\|_{P,2} \|A\|_{P,2} \leq C_1^2, \\ \left\| \mathbb{E}_P\left[R_A R_Y^T\right] \right\| &\leq \|A\|_{P,2} \|Y\|_{P,2} \leq C_1^2. \end{aligned}$$

In total, we hence have

$$\|b^\gamma\| \leq \frac{1}{c_4}\left(C_1^2 + |\gamma-1| \frac{C_1^4}{c_2} \right).$$ □

Lemma I.15. *Let $\gamma \geq 0$ The statistical rates r_N and λ_N introduced in Definition I.2 satisfy $r_N^4 \lesssim \delta_N$ and $\lambda_N \lesssim \frac{\delta_N}{\sqrt{N}}$.*

Proof of Lemma I.15. This proof is modified from Chernozhukov et al. [31]. First, we verify the bound on r_N. Let $S = (U, V, W, Z) \in \{A, X, Y\}^2 \times \{W\} \times$

$\{A, X, Y\}$, let $\eta = (m_U, m_V, m_Z) \in \mathcal{T}$, and let $b^0 \in \{b^\gamma, \beta_0, \mathbf{0}\}$. We have

$$\begin{aligned}
&\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0) \\
&= \big(U - m_U(W)\big)\Big(Z - m_Z(W) - \big(V - m_V(W)\big)^T b^0\Big)^T \\
&\quad -\big(U - m_U^0(W)\big)\Big(Z - m_Z^0(W) - \big(V - m_V^0(W)\big)^T b^0\Big)^T \\
&= \big(U - m_U^0(W)\big)\Big(m_Z^0(W) - m_Z(W) - \big(m_V^0(W) - m_V(W)\big)^T b^0\Big)^T \\
&\quad +\big(m_U^0(W) - m_U(W)\big)\Big(Z - m_Z^0(W) - \big(V - m_V^0(W)\big)^T b^0\Big)^T \\
&\quad +\big(m_U^0(W) - m_U(W)\big) \\
&\qquad \cdot\Big(m_Z^0(W) - m_Z(W) - \big(m_V^0(W) - m_V(W)\big)^T b^0\Big)^T.
\end{aligned}$$

By the triangle inequality and Hölder's inequality, we have

$$\begin{aligned}
&\mathbb{E}_P[\|\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0)\|] \\
&= \|\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0)\|_{P,1} \\
&\le \|U - m_U^0(W)\|_{P,2}\Big\|m_Z^0(W) - m_Z(W) - \big(m_V^0(W) - m_V(W)\big)^T b^0\Big\|_{P,2} \\
&\quad +\|m_U^0(W) - m_U(W)\|_{P,2}\Big\|Z - m_Z^0(W) - \big(V - m_V^0(W)\big)^T b^0\Big\|_{P,2} \\
&\quad +\|m_U^0(W) - m_U(W)\|_{P,2} \\
&\qquad \cdot\Big\|m_Z^0(W) - m_Z(W) - \big(m_V^0(W) - m_V(W)\big)^T b^0\Big\|_{P,2}.
\end{aligned}$$

Observe that $\|U - m_U^0(W)\|_{P,2} \le 2\|U\|_{P,2}$, and $\|V - m_V^0(W)\|_{P,2} \le 2\|V\|_{P,2}$, and $\|Z - m_Z^0(W)\|_{P,2} \le 2\|Z\|_{P,2}$ hold by Lemma I.7. We have $\|\eta - \eta^0\|_{P,2} \le \delta_N$ by Assumption I.5.5. Therefore, we obtain the upper bound

$$\begin{aligned}
&\mathbb{E}_P[\|\psi(S; b^0, \eta) - \psi(S; b^0, \eta^0)\|] \\
&\le 4\max\{1, \|b^0\|\}(\|U\|_{P,2} + \|V\|_{P,2} + \|Z\|_{P,2})\delta_N + 2\max\{1, \|b^0\|\}\delta_N^2 \\
&\lesssim \delta_N
\end{aligned}$$

by the triangle inequality, Lemma I.13, Lemma I.14, and Assumption I.5.2 and I.5.5. Because this upper bound is independent of η, we obtain our claimed bound on r_N^4.

Subsequently, we verify the bound on λ_N. Consider $S = (A, X, W, Y)$, denote by U either A or X, denote by Z either A or Y, and let $\varphi \in \{\psi, \widetilde{\psi}, \psi_2\}$, where we interpret $\psi_2(S; b, \eta) = \psi_2(S; \eta)$. We have

$$\begin{aligned}
&\partial_r^2\, \mathbb{E}_P\big[\psi\big(S; b^0, \eta^0 + r(\eta - \eta^0)\big)\big] \\
&= 2\,\mathbb{E}_P\Bigg[\big(m_U(W) - m_U^0(W)\big) \\
&\qquad\qquad \cdot\Big(m_Z(W) - m_Z^0(W) - \big(m_X(W) - m_X^0(W)\big)^T b^0\Big)^T\Bigg].
\end{aligned}$$

Due to the Cauchy–Schwarz inequality, we infer

$$\begin{aligned}&\big\|\partial_r^2\,\mathbb{E}_P\left[\psi\big(S;b^0,\eta^0+r(\eta-\eta^0)\big)\right]\big\|\\ &\leq 2\|m_U(W)-m_U^0(W)\|_{P,2}\\ &\qquad\cdot\big(\|m_Z(W)-m_Z^0(W)\|_{P,2}+\|m_X(W)-m_X^0(W)\|_{P,2}\|b^0\|\big)\\ &\leq 2\max\{1,\|b^0\|\}\|m_U(W)-m_U^0(W)\|_{P,2}\\ &\qquad\cdot\big(\|m_Z(W)-m_Z^0(W)\|_{P,2}+\|m_X(W)-m_X^0(W)\|_{P,2}\big)\\ &\lesssim \delta_N N^{-\frac{1}{2}}\end{aligned}$$

by Lemma I.13, Lemma I.14, and Assumption I.5.5. Consequently, we obtain our claimed bound on λ_N. □

Lemma I.16. *Let $\gamma \geq 0$. Let $k \in [K]$. Let furthermore $\varphi \in \{\psi, \widetilde{\psi}, \psi_2\}$ and $b^0 \in \{b^\gamma, \beta_0, \mathbf{0}\}$. We have*

$$\left\|\frac{1}{\sqrt{n}}\sum_{i\in I_k}\varphi(S_i;b^0,\hat{\eta}^{I_k^c})-\frac{1}{\sqrt{n}}\sum_{i\in I_k}\varphi(S_i;b^0,\eta^0)\right\|=O_P(\rho_N),$$

where $\rho_N = r_N + N^{\frac{1}{2}}\lambda_N$ is as in Definition I.4 and satisfies $\rho_N \lesssim \delta_N^{\frac{1}{4}}$, and where we interpret $\psi_2(S;b,\eta) = \psi_2(S;\eta)$.

Proof of Lemma I.16. This proof is modified from Chernozhukov et al. [31]. By the triangle inequality, we have

$$\begin{aligned}&\left\|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\varphi(S_i;b^0,\hat{\eta}^{I_k^c})-\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\varphi(S_i;b^0,\eta^0)\right\|\\ =&\left\|\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\big(\varphi(S_i;b^0,\hat{\eta}^{I_k^c})-\int\varphi(s;b^0,\hat{\eta}^{I_k^c})\mathrm{d}P(s)\big)\right.\\ &\qquad-\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\big(\varphi(S_i;b^0,\eta^0)-\int\varphi(s;b^0,\eta^0)\mathrm{d}P(s)\big)\\ &\qquad\left.+\sqrt{n}\textstyle\int\big(\varphi(s;b^0,\hat{\eta}^{I_k^c})-\varphi(s;b^0,\eta^0)\big)\mathrm{d}P(s)\right\|\\ \leq&\ \mathcal{I}_1+\sqrt{n}\mathcal{I}_2,\end{aligned}$$

where $\mathcal{I}_1 := \|M\|$ for

$$\begin{aligned}M:=&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\Big(\varphi(S_i;b^0,\hat{\eta}^{I_k^c})-\int\varphi(s;b^0,\hat{\eta}^{I_k^c})\mathrm{d}P(s)\Big)\\ &-\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\Big(\varphi(S_i;b^0,\eta^0)-\int\varphi(s;b^0,\eta^0)\mathrm{d}P(s)\Big),\end{aligned}$$

and where

$$\mathcal{I}_2:=\left\|\int\big(\varphi(s;b^0,\hat{\eta}^{I_k^c})-\varphi(s;b^0,\eta^0)\big)\mathrm{d}P(s)\right\|.$$

We bound the two terms $\mathcal{I}_1$ and $\mathcal{I}_2$ individually. First, we bound $\mathcal{I}_1$. Because the dimensions d and q are fixed, it is sufficient to bound one entry of the matrix M. Let l index the rows of M, and let t index the columns of M (we interpret vectors as matrices with one column). On the event $\mathcal{E}_N$ that holds with P-probability

$1-\Delta_N$, we have

$$
\begin{aligned}
&\mathbb{E}_P\left[\|M_{l,t}\|^2\middle|\{S_i\}_{i\in I_k^c}\right]\\
&=\tfrac{1}{n}\textstyle\sum_{i\in I_k}\mathbb{E}_P\left[|\varphi_{l,t}(S_i;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S_i;b^0,\eta^0)|^2\middle|\{S_i\}_{i\in I_k^c}\right]\\
&\quad+\tfrac{1}{n}\textstyle\sum_{i,j\in I_k,i\neq j}\mathbb{E}_P\left[\left(\varphi_{l,t}(S_i;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S_i;b^0,\eta^0)\right)\right.\\
&\qquad\qquad\left.\cdot\left(\varphi_{l,t}(S_j;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S_j;b^0,\eta^0)\right)\middle|\{S_i\}_{i\in I_k^c}\right]\\
&\quad-2\textstyle\sum_{i\in I_k}\mathbb{E}_P\left[\varphi_{l,t}(S_i;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S_i;b^0,\eta^0)\middle|\{S_i\}_{i\in I_k^c}\right]\\
&\qquad\cdot\mathbb{E}_P\left[\varphi_{l,t}(S;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S;b^0,\eta^0)\middle|\{S_i\}_{i\in I_k^c}\right]\\
&\quad+n\left|\mathbb{E}_P\left[\varphi_{l,t}(S;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S;b^0,\eta^0)\middle|\{S_i\}_{i\in I_k^c}\right]\right|^2\\
&=\mathbb{E}_P\left[|\varphi_{l,t}(S;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S;b^0,\eta^0)|^2\middle|\{S_i\}_{i\in I_k^c}\right]\\
&\quad+\left(\tfrac{n(n-1)}{n}-2n+n\right)\left|\mathbb{E}_P\left[\varphi_{l,t}(S;b^0,\hat{\eta}^{I_k^c})-\varphi_{l,t}(S;b^0,\eta^0)\middle|\{S_i\}_{i\in I_k^c}\right]\right|^2\\
&\le\sup_{\eta\in\mathcal{T}}\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^2\right].
\end{aligned}
\tag{25}
$$

Furthermore, for $\eta\in\mathcal{T}$, we have

$$
\begin{aligned}
&\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^2\right]\\
&\le\mathbb{E}_P[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|]\\
&\quad+\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^2\mathbb{1}_{\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\ge1}\right]
\end{aligned}
\tag{26}
$$

and

$$
\begin{aligned}
&\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^2\mathbb{1}_{\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\ge1}\right]\\
&\le\sqrt{\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^4\right]}\sqrt{P(\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\ge1)}
\end{aligned}
\tag{27}
$$

by Hölder's inequality. Observe that the term

$$
\sqrt{\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^4\right]}
\tag{28}
$$

is upper bounded by Assumption I.5.5, Lemma I.13, and Lemma I.14. By Markov's inequality, we have

$$
P(\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\ge1)\le\mathbb{E}_P[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|]\le r_N^4.
\tag{29}
$$

Therefore, we have $\mathbb{E}_P[\mathcal{I}_1^2|\{S_i\}_{i\in I_k^c}]\lesssim r_N^2$ due to (25)–(29). The statistical rate r_N satisfies $r_N\lesssim\delta_N^{\frac{1}{4}}$ by Lemma I.15. Thus, we infer $\mathcal{I}_1=O_P(r_N)$ by Lemma I.12. Subsequently, we bound $\mathcal{I}_2$. For $r\in[0,1]$, we introduce the function

$$
f_k(r):=\mathbb{E}_P\left[\varphi\left(S;b^0,\eta^0+r(\hat{\eta}^{I_k^c}-\eta^0)\right)\middle|\{S_i\}_{i\in I_k^c}\right]-\mathbb{E}_P[\varphi(S;b^0,\eta^0)].
$$

Observe that $\mathcal{I}_2=\|f_k(1)\|$ holds. We apply a Taylor expansion to this function and obtain

$$
f_k(1)=f_k(0)+f_k'(0)+\frac{1}{2}f_k''(\tilde{r})
$$

for some $\tilde{r}\in(0,1)$. We have

$$
f_k(0)=\mathbb{E}_P\left[\varphi(S;b^0,\eta^0)\middle|\{S_i\}_{i\in I_k^c}\right]-\mathbb{E}_P[\varphi(S;b^0,\eta^0)]=\mathbf{0}.
$$

Furthermore, the score φ satisfies the Neyman orthogonality property $f_k'(0) = \mathbf{0}$. The proof of this claim is analogous to the proof of Proposition 3.3 because the proof of Proposition 3.3 does neither depend on the underlying model of the random variables nor on the value of β. Furthermore, we have

$$f_k''(r) = 2\,\mathbb{E}_P\left[\big(m_U(W) - m_U^0(W)\big)\right.\\ \left.\cdot\Big(m_Z(W) - m_Z^0(W) - \big(m_X(W) - m_X^0(W)\big)^T b^0\Big)^T\right]$$

for $U \in \{A, X\}$ and $Z \in \{A, Y\}$. On the event $\mathcal{E}_N$ that holds with P-probability $1 - \Delta_N$, we have

$$\|f_k''(\tilde{r})\| \le \sup_{r\in(0,1)} \|f_k''(r)\| \lesssim \lambda_N.$$

We thus infer

$$\left\|\frac{1}{\sqrt{n}}\sum_{i\in I_k}\varphi(S_i; b^0, \hat{\eta}^{I_k^c}) - \frac{1}{\sqrt{n}}\sum_{i\in I_k}\varphi(S_i; b^0, \eta^0)\right\| \le \mathcal{I}_1 + \sqrt{n}\mathcal{I}_2 = O_P(r_N + N^{\frac{1}{2}}\lambda_N).$$

Because $r_N \lesssim \delta_N^{\frac{1}{4}}$ and $\lambda_N \lesssim \frac{\delta_N}{\sqrt{N}}$ hold by Lemma I.15 and because $\{\delta_N\}_{N\ge K}$ converges to 0 by Assumption I.5, we furthermore have

$$\rho_N = r_N + N^{\frac{1}{2}}\lambda_N \lesssim \delta_N^{\frac{1}{4}}.$$ □

Lemma I.17. *Let $k \in [K]$. Let furthermore $U, V \in \{A, X\}$ and $S = (U, V, W, Y)$. Let $\varphi \in \{\psi_1, \psi_2, \psi_3\}$. We have*

$$\frac{1}{n}\sum_{i\in I_k}\varphi(S_i; \hat{\eta}^{I_k^c}) = \mathbb{E}_P[\varphi(S; \eta^0)] + O_P\big(N^{-\frac{1}{2}}(1 + \rho_N)\big).$$

Proof of Lemma I.17. Consider the decomposition

$$\begin{aligned}&\tfrac{1}{n}\textstyle\sum_{i\in I_k}\varphi(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\varphi(S; \eta^0)]\\ =\ &\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\varphi(S_i; \hat{\eta}^{I_k^c}) - \varphi(S_i; \eta^0)\big) + \tfrac{1}{n}\sum_{i\in I_k}\big(\varphi(S_i; \eta^0) - \mathbb{E}_P[\varphi(S; \eta^0)]\big).\end{aligned}$$

Because of Lemma I.16, the term $\frac{1}{n}\sum_{i\in I_k}\big(\varphi(S_i; \hat{\eta}^{I_k^c}) - \varphi(S_i; \eta^0)\big)$ is of order $O_P(N^{-\frac{1}{2}}\rho_N)$. The term $\frac{1}{n}\sum_{i\in I_k}\big(\varphi(S_i; \eta^0) - \mathbb{E}_P[\varphi(S; \eta^0)]\big)$ is of order $O_P(N^{-\frac{1}{2}})$ due to the Lindeberg–Feller CLT and the Cramer–Wold device. Thus, we deduce the statement. □

Definition I.18. *We denote by $\boldsymbol{A}^{I_k}$ the row-wise concatenation of the observations A_i for $i \in I_k$. We denote similarly by $\boldsymbol{X}^{I_k}$, $\boldsymbol{W}^{I_k}$, $\boldsymbol{Y}^{I_k}$, $\boldsymbol{A}^{I_k^c}$, $\boldsymbol{X}^{I_k^c}$, $\boldsymbol{W}^{I_k^c}$, and $\boldsymbol{Y}^{I_k^c}$ the row-wise concatenations of the respective observations.*

Proof of Theorem 3.1. This proof is based on Chernozhukov et al. [31]. We show the stronger statement

$$\sqrt{N}\sigma^{-1}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\psi}(S_i; \beta_0, \eta^0) + O_P(\rho_N) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d\times d}) \quad (N \to \infty), \tag{30}$$

where $\hat{\beta}$ denotes the DML1 estimator $\hat{\beta}^{\mathrm{DML1}}$ or the DML2 estimator $\hat{\beta}^{\mathrm{DML2}}$, and where the rate ρ_N is specified in Definition I.4, and we show that this statement holds uniformly over laws P. We first consider $\hat{\beta}^{\mathrm{DML2}}$. It suffices to show that (30) holds uniformly over $P \in \mathcal{P}_N$. Fix a sequence $\{P_N\}_{N\geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, it suffices to show

$$\begin{aligned}&\sqrt{N}\sigma^{-1}(\hat{\beta}^{\mathrm{DML2}} - \beta_0)\\ &= \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N \overline{\psi}(S_i;\beta_0,\eta^0) + O_{P_N}(\rho_N)\\ &\xrightarrow{d} \mathcal{N}(0,\mathbb{1}_{d\times d}) \quad (N\to\infty).\end{aligned}$$

We have

$$\begin{aligned}&\hat{\beta}^{\mathrm{DML2}}\\
&= \Big(\tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \Pi_{\hat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad \cdot \tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \Pi_{\hat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&= \Bigg(\tfrac{1}{K}\textstyle\sum_{k=1}^K \tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad \cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Bigg)^{-1}\\
&\quad \cdot\tfrac{1}{K}\textstyle\sum_{k=1}^K \tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad \cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\end{aligned} \tag{31}$$

by (7). By Lemma I.17, we have

$$\begin{aligned}&\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\ &= \mathbb{E}_{P_N}\Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T\Big] + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\end{aligned} \tag{32}$$

and

$$\begin{aligned}&\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\ &= \mathbb{E}_{P_N}\Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T\Big] + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big).\end{aligned} \tag{33}$$

Recall the matrix J_0 introduced in Definition I.1. By Weyl's inequality and

Slutsky's theorem, combining Equations (31)–(33) gives

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML2}}-\beta_0)\\
&=\Bigg(\bigg(\mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]^{-1}\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(X-m_X^0(W)\big)^T\Big]\bigg)^{-1}\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Bigg)\\
&\qquad\cdot\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\tfrac{1}{\sqrt{n}}\Big(\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\qquad-\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big)\\
&=\big(J_0+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\big)\\
&\qquad\cdot\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\tfrac{1}{\sqrt{n}}\bigg(\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad\qquad\cdot\Big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})-\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big)\bigg)
\end{aligned}\tag{34}$$

because K is a constant independent of N and because $N = nK$ holds. Recall the linear score ψ in (11). We have

$$\sqrt{N}(\hat{\beta}^{\mathrm{DML2}}-\beta_0)=\Big(J_0+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\frac{1}{\sqrt{K}}\sum_{k=1}^K\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\hat{\eta}^{I_k^c}).\tag{35}$$

Let $k\in[K]$. By Lemma I.16, we have

$$\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\hat{\eta}^{I_k^c})=\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\eta^0)+O_{P_N}(\rho_N).\tag{36}$$

We combine (35) and (36) to obtain

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML2}}-\beta_0)\\
&=\Big(J_0+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\hat{\eta}^{I_k^c})\\
&=\Big(J_0+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\Big(\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\eta^0)+O_{P_N}(\rho_N)\Big).
\end{aligned}$$

Recall that we have $N = nK$, that K is a constant independent of N, that the sets I_k for $k\in[K]$ form a partition of $[N]$, that $\rho_N\lesssim\delta_N^{\frac{1}{4}}$ by Lemma I.16, and that δ_N converges to 0 as $N\to\infty$ and $\delta_N^{\frac{1}{4}}\geq N^{-\frac{1}{2}}$ holds by Assumption I.5.

Thus, we have

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML2}} - \beta_0) \\
&= \Big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \Big(\tfrac{1}{\sqrt{n}} \sum_{i \in I_k} \psi(S_i; \beta_0, \eta^0) + O_{P_N}(\rho_N)\Big) \\
&= \Big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N \big(\psi(S_i; \beta_0, \eta^0) + O_{P_N}(\rho_N)\big) \\
&= J_0 \cdot \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N \psi(S_i; \beta_0, \eta^0) + O_{P_N}(\rho_N).
\end{aligned}$$

We have $\mathbb{E}_{P_N}[\psi(S; \beta_0, \eta^0)] = \mathbf{0}$ due to the identifiability condition (5). Therefore, we conclude the proof concerning the DML2 method due to the Lindeberg–Feller CLT and the Cramer–Wold device.

Subsequently, we consider the DML1 method. It suffices to show that (30) holds uniformly over $P \in \mathcal{P}_N$. Fix a sequence $\{P_N\}_{N \geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, it suffices to show

$$\begin{aligned}
&\sqrt{N}\sigma^{-1}(\hat{\beta}^{\mathrm{DML1}} - \beta_0) \\
&= \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N \overline{\psi}(S_i; \beta_0, \eta^0) + O_{P_N}(\rho_N) \\
&\stackrel{d}{\to} \mathcal{N}(0, \mathbb{1}_{d \times d}) \quad (N \to \infty).
\end{aligned}$$

We have

$$\begin{aligned}
&\hat{\beta}^{I_k} \\
&= \Big(\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1} \\
&\quad \cdot \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&= \Bigg(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad \cdot \Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1} \\
&\qquad \cdot \tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Bigg)^{-1} \\
&\quad \cdot \tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad \cdot \Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1} \\
&\qquad \cdot \tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)
\end{aligned} \tag{37}$$

by (19). Due to Weyl's inequality and Slutsky's theorem, (32), (33), and (37), we obtain

$$
\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML1}} - \beta_0) \\
&= \Bigg(\bigg(\mathbb{E}_{P_N} \Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T \Big] \\
&\qquad \cdot \mathbb{E}_{P_N} \Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T \Big]^{-1} \\
&\qquad \cdot \mathbb{E}_{P_N} \Big[\big(A - m_A^0(W)\big)\big(X - m_X^0(W)\big)^T \Big] \bigg)^{-1} \\
&\quad \cdot \mathbb{E}_{P_N} \Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T \Big] \\
&\quad \cdot \mathbb{E}_{P_N} \Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T \Big]^{-1} + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \Bigg) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \Big(\tfrac{1}{\sqrt{n}} \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad\qquad - \tfrac{1}{\sqrt{n}} \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0 \Big) \\
&= \Big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \bigg(\tfrac{1}{\sqrt{n}} \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \\
&\qquad\qquad \cdot \Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0 \Big) \bigg).
\end{aligned} \tag{38}
$$

Observe that the expression for $\sqrt{N}(\hat{\beta}^{\mathrm{DML1}} - \beta_0)$ given in (38) coincides with the expression for $\sqrt{N}(\hat{\beta}^{\mathrm{DML2}} - \beta_0)$ given in (34). Thus, the asymptotic analysis of $\sqrt{N}(\hat{\beta}^{\mathrm{DML1}} - \beta_0)$ coincides with the asymptotic analysis of $\sqrt{N}(\hat{\beta}^{\mathrm{DML2}} - \beta_0)$ presented above. □

Lemma I.19. *Let $\gamma \geq 0$. Let $p > 4$ be the p from Assumption I.5, let $b^0 \in \{\beta_0, b^\gamma, \mathbf{0}\}$, and let $S = (U, V, Z) \in \{A, X, Y\}^2 \times \{W\} \times \{A, X, Y\}$. There exists a finite real constant C_5 satisfying*

$$
\sup_{\eta \in \mathcal{T}} \mathbb{E}_P \Big[\|\psi(S; b^0, \eta)\|^{\frac{p}{2}} \Big]^{\frac{2}{p}} \leq C_5.
$$

Proof of Lemma I.19. Let $\eta = (m_U, m_V, m_Z) \in \mathcal{T}$. By Hölder's inequality and the triangle inequality, we have

$$
\begin{aligned}
&\mathbb{E}_P \Big[\|\psi(S; b^0, \eta)\|^{\frac{p}{2}} \Big]^{\frac{2}{p}} \\
&= \|(U - m_U(W))\big(Z - m_Z(W) - (V - m_V(W))^T b^0\big)\|_{P, \frac{2}{p}} \\
&\leq \big(\|U - m_U^0(W)\|_{P,p} + \|m_U^0(W) - m_U(W)\|_{P,p} \big) \\
&\quad \cdot \big(\|Z - m_Z^0(W)\|_{P,p} + \|(V - m_V^0(W))^T b^0\|_{P,p} \\
&\qquad + \|m_Z^0(W) - m_Z(W)\|_{P,p} + \|(m_V^0(W) - m_V(W))^T b^0\|_{P,p} \big).
\end{aligned} \tag{39}
$$

By the Cauchy–Schwarz inequality, we have

$$
\Big\| \big(V - m_V^0(W)\big)^T b^0 \Big\|_{P,p} \leq \mathbb{E}_P \big[\|V - m_V^0(W)\|^p \|b^0\|^p \big]^{\frac{1}{p}} = \|b^0\| \|V - m_V^0(W)\|_{P,p} \tag{40}
$$

and analogously

$$\left\| \left(m_V^0(W) - m_V(W)\right)^T b^0 \right\|_{P,p} \leq \|b^0\| \|m_V^0(W) - m_V(W)\|_{P,p}. \quad (41)$$

Hence, we infer

$$\mathbb{E}_P\left[\|\psi(S;b^0,\eta)\|^{\frac{p}{2}}\right]^{\frac{2}{p}} \leq (\|U\|_{P,p} + C_2)(\|Z\|_{P,p} + \|V\|_{P,p} + 2C_2)\max\{1, \|b^0\|\} \quad (42)$$

by (39), (40), (41), Lemma I.7, and Assumption I.5.5. By Lemma I.13, there exists a finite real constant C_3 that satisfies $\|\beta_0\| \leq C_3$. By Lemma I.14, there exists a finite real constant C_4 that satisfies $\|b^\gamma\| \leq C_4$. These two bounds lead to $\|b^0\| \leq \max\{C_3, C_4\}$. By Assumption I.5.2, we have

$$\max\{\|U\|_{P,p}, \|V\|_{P,p}, \|Z\|_{P,p}\} \leq \|U\|_{P,p} + \|V\|_{P,p} + \|Z\|_{P,p} \leq 3C_1.$$

Due to (42), we therefore have

$$\mathbb{E}_P\left[\|\psi(S;b^0,\eta)\|^{\frac{p}{2}}\right]^{\frac{2}{p}} \leq (3C_1 + C_2)(6C_1 + 2C_2)\max\{1, C_3, C_4\}. \qquad \square$$

Lemma I.20. *Let $\gamma \geq 0$, and let p be as in Assumption I.5. Let the indices $k \in [K]$ and $(j,l,t,r) \in [L_1] \times [L_2] \times [L_3] \times [L_4]$, where L_1, L_2, L_3, and L_4 are natural numbers representing the intended dimensions. Let $\hat{b} \in \{\hat{\beta}^{DML1}, \hat{\beta}^{DML2}, \hat{b}^{\gamma,DML1}, \hat{b}^{\gamma,DML2}\}$, and consider the corresponding true but unknown underlying parameter vector $b^0 \in \{\beta_0, b^\gamma\}$. Consider the corresponding score function combinations*

$$\begin{aligned}
\hat{\psi}^A(\cdot) &\in \{\widetilde{\psi}_j(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi_j(\cdot;\hat{b},\hat{\eta}^{I_k^c}), (\psi_1(\cdot;\hat{\eta}^{I_k^c}))_{j,l}, (\psi_2(\cdot;\hat{\eta}^{I_k^c}))_{j,l}\},\\
\hat{\psi}^A_{full}(\cdot) &\in \{\widetilde{\psi}(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi_1(\cdot;\hat{\eta}^{I_k^c}), \psi_2(\cdot;\hat{\eta}^{I_k^c})\},\\
\hat{\psi}^B(\cdot) &\in \{\widetilde{\psi}_t(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi_t(\cdot;\hat{b},\hat{\eta}^{I_k^c}), (\psi_1(\cdot;\hat{\eta}^{I_k^c}))_{t,r}, (\psi_2(\cdot;\hat{\eta}^{I_k^c}))_{t,r}\},\\
\hat{\psi}^B_{full}(\cdot) &\in \{\widetilde{\psi}(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi(\cdot;\hat{b},\hat{\eta}^{I_k^c}), \psi_1(\cdot;\hat{\eta}^{I_k^c}), \psi_2(\cdot;\hat{\eta}^{I_k^c})\}
\end{aligned}$$

and their respective nonestimated quantity

$$\begin{aligned}
\psi^A(\cdot) &\in \{\widetilde{\psi}_j(\cdot;b^0,\eta^0), \psi_j(\cdot;b^0,\eta^0), (\psi_1(\cdot;\eta^0))_{j,l}, (\psi_2(\cdot;\eta^0))_{j,l}\},\\
\psi^A_{full}(\cdot) &\in \{\widetilde{\psi}(\cdot;b^0,\eta^0), \psi(\cdot;b^0,\eta^0), \psi_1(\cdot;\eta^0), \psi_2(\cdot;\eta^0)\},\\
\psi^B(\cdot) &\in \{\widetilde{\psi}_t(\cdot;b^0,\eta^0), \psi_t(\cdot;b^0,\eta^0), (\psi_1(\cdot;\eta^0))_{t,r}, (\psi_2(\cdot;\eta^0))_{t,r}\},\\
\psi^B_{full}(\cdot) &\in \{\widetilde{\psi}(\cdot;b^0,\eta^0), \psi(\cdot;b^0,\eta^0), \psi_1(\cdot;\eta^0), \psi_2(\cdot;\eta^0)\}.
\end{aligned}$$

Then we have

$$\mathcal{I}_k := \left| \frac{1}{n}\sum_{i\in I_k} \hat{\psi}^A(S_i)\hat{\psi}^B(S_i) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right]\right| = O_P(\tilde{\rho}_N),$$

where $\tilde{\rho}_N = N^{\max\left\{\frac{4}{p}-1, -\frac{1}{2}\right\}} + r_N$ is as in Definition I.4.

Proof of Lemma I.20. This proof is modified from Chernozhukov et al. [31]. By the triangle inequality, we have

$$\mathcal{I}_k \leq \mathcal{I}_{k,A} + \mathcal{I}_{k,B},$$

where

$$\mathcal{I}_{k,A} := \left| \frac{1}{n} \sum_{i \in I_k} \hat{\psi}^A(S_i)\hat{\psi}^B(S_i) - \frac{1}{n} \sum_{i \in I_k} \psi^A(S_i)\psi^B(S_i) \right|$$

and

$$\mathcal{I}_{k,B} := \left| \frac{1}{n} \sum_{i \in I_k} \psi^A(S_i)\psi^B(S_i) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right] \right|.$$

Subsequently, we bound the two terms $\mathcal{I}_{k,A}$ and $\mathcal{I}_{k,B}$ individually. First, we bound $\mathcal{I}_{k,B}$. We consider the case $p \leq 8$. The von Bahr–Esseen inequality I [37, p. 650] states that for $1 \leq u \leq 2$ and for independent, real-valued, and mean 0 variables $Z_1, \ldots, Z_n$, we have

$$\mathbb{E}\left[\left| \sum_{i=1}^n Z_i \right|^u \right] \leq \left(2 - \frac{1}{n}\right) \sum_{i=1}^n \mathbb{E}[|Z_i|^u].$$

The individual summands $\psi^A(S_i)\psi^B(S_i) - \mathbb{E}_P[\psi^A(S)\psi^B(S)]$ for $i \in I_k$ are independent and have mean 0. Therefore,

$$\begin{aligned}
&\mathbb{E}_P\left[\mathcal{I}_{k,B}^{\frac{p}{4}}\right] \\
&= \left(\tfrac{1}{n}\right)^{\frac{p}{4}} \mathbb{E}_P\left[\left| \textstyle\sum_{i \in I_k} \left(\psi^A(S_i)\psi^B(S_i) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right]\right) \right|^{\frac{p}{4}} \right] \\
&\leq \left(\tfrac{1}{n}\right)^{-1+\frac{p}{4}} \left(2 - \tfrac{1}{n}\right) \tfrac{1}{n} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\left| \psi^A(S_i)\psi^B(S_i) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right] \right|^{\frac{p}{4}} \right] \\
&= \left(\tfrac{1}{n}\right)^{-1+\frac{p}{4}} \left(2 - \tfrac{1}{n}\right) \mathbb{E}_P\left[\left| \psi^A(S)\psi^B(S) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right] \right|^{\frac{p}{4}} \right]
\end{aligned}$$

follows due to the von Bahr–Esseen inequality I because $1 < \frac{p}{4} \leq 2$ holds. By Hölder's inequality, we have

$$\begin{aligned}
&\left(\mathbb{E}_P\left[\left|\psi^A(S)\right|^{\frac{p}{4}} \left|\psi^B(S)\right|^{\frac{p}{4}} \right] \right)^{\frac{p}{4}} \\
&\leq \mathbb{E}_P\left[\left|\psi^A(S)\right|^{\frac{p}{2}} \right]^{\frac{2}{p}} \mathbb{E}_P\left[\left|\psi^B(S; b^\gamma, \eta^0)\right|^{\frac{p}{2}} \right]^{\frac{2}{p}} \\
&\leq \left\|\psi^A_{\text{full}}(S)\right\|_{P,\frac{p}{2}} \left\|\psi^B_{\text{full}}(S)\right\|_{P,\frac{p}{2}}.
\end{aligned}$$

The terms $\|\psi(S; b^0, \eta^0)\|_{P,\frac{p}{2}}$, $\|\widetilde{\psi}(S; b^0, \eta^0)\|_{P,\frac{p}{2}}$, $\|\psi_1(S; \eta)\|_{P,\frac{p}{2}}$, and $\|\psi_2(S; \eta)\|_{P,\frac{p}{2}}$ are upper bounded by the finite real constant C_5 by Lemma I.19. Thus, we have $\mathcal{I}_{k,B} = O_P(N^{\frac{p}{4}-1})$ by Lemma I.12 because we have

$$\begin{aligned}
&\mathbb{E}_P\left[\left| \psi^A(S)\psi^B(S) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right] \right|^{\frac{p}{4}} \right]^{\frac{4}{p}} \\
&= \|\psi^A(S)\psi^B(S) - \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right]\|_{P,\frac{p}{4}} \\
&\leq \|\psi^A(S)\psi^B(S)\|_{P,\frac{p}{4}} + \mathbb{E}_P\left[|\psi^A(S)\psi^B(S)|\right] \\
&\leq 2\|\psi^A(S)\psi^B(S)\|_{P,\frac{p}{4}}
\end{aligned}$$

by the triangle inequality, Hölder's inequality, and due to $\frac{p}{4} > 1$.

Next, consider the case $p > 8$. Observe that

$$\begin{aligned} &\mathbb{E}_P\left[\left(\tfrac{1}{n}\sum_{i\in I_k}\psi^A(S_i)\psi^B(S_i)\right)^2\right] \\ &= \tfrac{1}{n}\,\mathbb{E}_P\left[\left(\psi^A(S)\right)^2\left(\psi^B(S)\right)^2\right] + \tfrac{n(n-1)}{n^2}\,\mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right]^2 \end{aligned}$$

holds because the data sample is iid. Thus, we infer

$$\begin{aligned} &\mathbb{E}_P[\mathcal{I}_{k,B}^2] \\ &= \mathbb{E}_P\left[\left(\tfrac{1}{n}\sum_{i\in I_k}\psi^A(S_i)\psi^B(S_i)\right)^2\right] + \mathbb{E}_P\left[\psi^A(S)\psi^B(S)\right]^2 \\ &\quad -2\,\mathbb{E}_P\left[\tfrac{1}{n}\sum_{i\in I_k}\psi^A(S_i)\psi^B(S_i)\right]\mathbb{E}_P[\psi^A(S)\psi^B(S)] \\ &\le \tfrac{1}{n}\,\mathbb{E}_P\left[(\psi^A(S))^2(\psi^B(S))^2\right]. \end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} &\tfrac{1}{n}\,\mathbb{E}_P\left[\left(\psi^A(S)\right)^2\left(\psi^B(S)\right)^2\right] \\ &\le \tfrac{1}{n}\sqrt{\mathbb{E}_P\left[\left(\psi^A(S)\right)^4\right]\mathbb{E}_P\left[\left(\psi^B(S)\right)^4\right]} \\ &\le \tfrac{1}{n}\left\|\psi^A_{\text{full}}(S)\right\|^2_{P,4}\left\|\psi^B_{\text{full}}(S)\right\|^2_{P,4}. \end{aligned}$$

The terms $\|\psi(S;b^0,\eta^0)\|_{P,4}$, $\|\widetilde{\psi}(S;b^0,\eta^0)\|_{P,4}$, $\|\psi_1(S;\eta)\|_{P,4}$, and $\|\psi_2(S;\eta)\|_{P,4}$ are upper bounded by C_5 by Lemma I.19. Thus, we have

$$\mathbb{E}_P[\mathcal{I}_{k,B}^2] \le \frac{1}{n}\left\|\psi^A_{\text{full}}(S)\right\|^2_{P,4}\left\|\psi^B_{\text{full}}(S)\right\|^2_{P,4} \le \frac{1}{n}(4C_5)^4.$$

We hence infer $\mathcal{I}_{k,B} = O_P(N^{-\frac{1}{2}})$ by Lemma I.12.

Second, we bound the term $\mathcal{I}_{k,A}$. For any real numbers a_1, a_2, b_1, and b_2 such that real numbers c and d exist that satisfy $\max\{|b_1|,|b_2|\} \le c$ and $\max\{|a_1 - b_1|,|a_2-b_2|\} \le d$, we have $|a_1a_2 - b_1b_2| \le 2d(c+d)$. Indeed, we have

$$\begin{aligned} &|a_1a_2 - b_1b_2| \\ &\le |a_1-b_1|\cdot|a_2-b_2| + |b_1|\cdot|a_2-b_2| + |a_1-b_1|\cdot|b_2| \\ &\le d^2 + cd + dc \\ &\le 2d(c+d) \end{aligned}$$

by the triangle inequality.

We apply this observation together with the triangle inequality and the

Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
&\mathcal{I}_{k,A} \\
&\leq \tfrac{1}{n}\textstyle\sum_{i\in I_k}\big|\hat{\psi}^A(S_i)\hat{\psi}^B(S_i)-\psi^A(S_i)\psi^B(S_i)\big| \\
&\leq \tfrac{2}{n}\textstyle\sum_{i\in I_k}\max\big\{\big|\hat{\psi}^A(S_i)-\psi^A(S_i)\big|,\big|\hat{\psi}^B(S_i)-\psi^B(S_i)\big|\big\} \\
&\qquad\cdot\Big(\max\big\{\big|\psi^A(S_i)\big|,\big|\psi^B(S_i)\big|\big\} \\
&\qquad\qquad+\max\big\{\big|\hat{\psi}^A(S_i)-\psi^A(S_i)\big|,\big|\hat{\psi}^B(S_i)-\psi^B(S_i)\big|\big\}\Big) \\
&\leq 2\Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\max\Big\{\big|\hat{\psi}^A(S_i)-\psi^A(S_i)\big|^2,\big|\hat{\psi}^B(S_i)-\psi^B(S_i)\big|^2\Big\}\Big)^{\frac{1}{2}} \\
&\qquad\cdot\Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\max\big\{\big|\psi^A(S_i)\big|,\big|\psi^B(S_i)\big|\big\} \\
&\qquad\qquad\qquad+\max\big\{\big|\hat{\psi}^A(S_i)-\psi^A(S_i)\big|,\big|\hat{\psi}^B(S_i)-\psi^B(S_i)\big|\big\}\big)^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

By the triangle inequality, we hence have

$$\mathcal{I}_{k,A}^2 \leq 4R_{N,k}\bigg(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\Big(\big\|\psi_{\mathrm{full}}^A(S_i)\big\|^2+\big\|\psi_{\mathrm{full}}^B(S_i)\big\|^2\Big)+R_{N,k}\bigg) \tag{43}$$

by Lemma I.11, where

$$R_{N,k} := \frac{1}{n}\sum_{i\in I_k}\Big(\big\|\hat{\psi}_{\mathrm{full}}^A(S_i)-\psi_{\mathrm{full}}^A(S_i)\big\|^2+\big\|\hat{\psi}_{\mathrm{full}}^B(S_i)-\psi_{\mathrm{full}}^B(S_i)\big\|^2\Big).$$

Note that we have

$$\frac{1}{n}\sum_{i\in I_k}\Big(\big\|\psi_{\mathrm{full}}^A(S_i)\big\|^2+\big\|\psi_{\mathrm{full}}^B(S_i)\big\|^2\Big)=O_P(1)$$

by Markov's inequality because the terms $\|\psi(S;b^0,\eta^0)\|_{P,4}$, $\|\widetilde{\psi}(S;b^0,\eta^0)\|_{P,4}$, $\|\psi_1(S;\eta)\|_{P,4}$, and $\|\psi_2(S;\eta)\|_{P,4}$ are upper bounded by C_5 by Lemma I.19. Thus, it suffices to bound the term $R_{N,k}$. To do this, we need to bound the four terms

$$\frac{1}{n}\sum_{i\in I_k}\|\psi(S_i;\hat{b},\hat{\eta}^{I_k^c})-\psi(S_i;b^0,\eta^0)\|^2, \tag{44}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\widetilde{\psi}(S_i;\hat{b},\hat{\eta}^{I_k^c})-\widetilde{\psi}(S_i;b^0,\eta^0)\|^2, \tag{45}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_1(S_i;\hat{\eta}^{I_k^c})-\psi_1(S_i;\eta^0)\|^2, \tag{46}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_2(S_i;\hat{\eta}^{I_k^c})-\psi_2(S_i;\eta^0)\|^2. \tag{47}$$

First, we bound the two terms (44) and (45) simultaneously. Consider the random variable $U\in\{A,X\}$ and the quadruple $S=(U,X,W,Y)$. Because the score ψ is linear in β, these two terms are upper bounded by

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|-\psi^a(S_i;\hat\eta^{I_k^c})(\hat b - b^0) + \psi(S_i;b^0,\hat\eta^{I_k^c}) - \psi(S_i;b^0,\eta^0)\|^2 \\ &\le \tfrac{2}{n}\textstyle\sum_{i\in I_k}\|\psi^a(S_i;\hat\eta^{I_k^c})(\hat b - b^0)\|^2 + \tfrac{2}{n}\sum_{i\in I_k}\|\psi(S_i;b^0,\hat\eta^{I_k^c}) - \psi(S_i;b^0,\eta^0)\|^2 \end{aligned} \tag{48}$$

due to the triangle inequality and Lemma I.11. Subsequently, we verify that

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat\eta^{I_k^c})\|^2 = O_P(1)$$

holds. Indeed, we have

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\psi^a(S_i;\hat\eta^{I_k^c})\|^2 \\ &= \tfrac{1}{n}\textstyle\sum_{i\in I_k}\left\|\big(U_i - \hat m_U^{I_k^c}(W_i)\big)\big(X_i - \hat m_X^{I_k^c}(W_i)\big)^T\right\|^2 \\ &\le \sqrt{\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|U_i - \hat m_U^{I_k^c}(W_i)\|^4}\sqrt{\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|X_i - \hat m_X^{I_k^c}(W_i)\|^4} \end{aligned} \tag{49}$$

by the Cauchy–Schwarz inequality. We have

$$\left(\frac{1}{n}\sum_{i\in I_k}\|U_i - m_U^0(W_i)\|^4\right)^{\frac{1}{4}} = O_P(1) \tag{50}$$

by Markov's inequality because the term $\mathbb{E}_P[\|U - m_U^0(W)\|^4]$ is upper bounded by Lemma I.7 and Assumption I.5.2. On the event $\mathcal{E}_N$ that holds with P-probability $1-\Delta_N$, we have

$$\begin{aligned} &\mathbb{E}_P\left[\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\eta^0(W_i) - \hat\eta^{I_k^c}(W_i)\|^4 \middle| \{S_i\}_{i\in I_k^c}\right] \\ &= \mathbb{E}_P\left[\|\eta^0(W) - \hat\eta^{I_k^c}(W)\|^4 | \{S_i\}_{i\in I_k^c}\right] \\ &\le C_2^4 \end{aligned} \tag{51}$$

by Assumption I.5.5. We hence have $\frac{1}{n}\sum_{i\in I_k}\|\eta^0(W_i) - \hat\eta^{I_k^c}(W_i)\| = O_P(1)$ by Lemma I.12. Let us denote by $\|\cdot\|_{P_{I_k},p}$ the L^p-norm with the empirical measure on the data indexed by I_k. On the event $\mathcal{E}_N$ that holds with P-probability $1-\Delta_N$, we have

$$\begin{aligned} &\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|U_i - \hat m_U^{I_k^c}(W_i)\|^4 \\ &= \|U - \hat m_U^{I_k^c}(W)\|_{P_{I_k},4}^4 \\ &\le \big(\|U - m_U^0(W)\|_{P_{I_k},4} + \|m_U^0(W) - \hat m_U^{I_k^c}(W)\|_{P_{I_k},4}\big)^4 \\ &\le \big(\|U - m_U^0(W)\|_{P_{I_k},4} + \|\eta^0(W) - \hat\eta^{I_k^c}(W)\|_{P_{I_k},4}\big)^4 \\ &= O_P(1) \end{aligned} \tag{52}$$

by the triangle inequality, (50), and (51). Analogous arguments lead to

$$\frac{1}{n}\sum_{i\in I_k}\|X_i - \hat m_X^{I_k^c}(W_i)\|^4 = O_P(1). \tag{53}$$

We combine (49), (52), and (53) to obtain

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})\|^2 = O_P(1). \tag{54}$$

Because $\|\hat{b}-b^0\|^2 = O_P(N^{-1})$ holds by Theorem 3.1 and Theorem 4.1, we can bound the first summand in (48) by

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})(\hat{b}-b^0)\|^2 = O_P(1)O_P(N^{-1}) = O_P(N^{-1}) \tag{55}$$

due to the Cauchy–Schwarz inequality and (54). On the event $\mathcal{E}_N$ that holds with P-probability $1-\Delta_N$, the conditional expectation given $\{S_i\}_{i\in I_k^c}$ of the second summand in (48) is equal to

$$\begin{aligned}
&\mathbb{E}_P\left[\tfrac{2}{n}\textstyle\sum_{i\in I_k}\|\psi(S_i;b^0,\hat{\eta}^{I_k^c})-\psi(S_i;b^0,\eta^0)\|^2\Big|\{S_i\}_{i\in I_k^c}\right]\\
&= 2\,\mathbb{E}_P\left[\|\psi(S;b^0,\hat{\eta}^{I_k^c})-\psi(S;b^0,\eta^0)\|^2\big|\{S_i\}_{i\in I_k^c}\right]\\
&\le 2\sup\nolimits_{\eta\in\mathcal{T}}\mathbb{E}_P\left[\|\psi(S;b^0,\eta)-\psi(S;b^0,\eta^0)\|^2\right]\\
&\lesssim r_N^2
\end{aligned}$$

due to arguments that are analogous to (25)–(29) presented in the proof of Lemma I.16. Because the event $\mathcal{E}_N$ holds with P-probability $1-\Delta_N = 1-o(1)$, we infer

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})(\hat{b}-b^0)+\psi(S_i;b^0,\hat{\eta}^{I_k^c})-\psi(S_i;b^0,\eta^0)\|^2 = O_P(N^{-1}+r_N^2)$$

by Lemma I.12. Next, we bound the two terms given in (46) and (47). We first consider the term given in (46). On the event $\mathcal{E}_N$, we have

$$\begin{aligned}
&\mathbb{E}_P\left[\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\psi_1(S_i;\hat{\eta}^{I_k^c})-\psi_1(S_i;\eta^0)\|^2\Big|\{S_i\}_{i\in I_k^c}\right]\\
&= \mathbb{E}_P\left[\|\psi_1(S;\hat{\eta}^{I_k^c})-\psi_1(S;\eta^0)\|^2\big|\{S_i\}_{i\in I_k^c}\right]\\
&\le \sup\nolimits_{\eta\in\mathcal{T}}\mathbb{E}_P\left[\|\psi_1(S;\eta)-\psi_1(S;\eta^0)\|^2\right]\\
&\lesssim r_N^2
\end{aligned}$$

due to arguments that are analogous to (25)–(29) presented in the proof of Lemma I.16. Because the event $\mathcal{E}_N$ holds with probability $1-\Delta_N = 1-o(1)$, we infer

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_1(S_i;\hat{\eta}^{I_k^c})-\psi_1(S_i;\eta^0)\|^2 = O_P(r_N^2)$$

by Lemma I.12. On the event $\mathcal{E}_N$, the conditional expectation given $\{S_i\}_{i\in I_k^c}$ of the term (47) is given by

$$\begin{aligned}
&\mathbb{E}_P\left[\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\psi_2(S_i;\hat{\eta}^{I_k^c})-\psi_2(S_i;\eta^0)\|^2\Big|\{S_i\}_{i\in I_k^c}\right]\\
&= \mathbb{E}_P\left[\|\psi_2(S;\hat{\eta}^{I_k^c})-\psi_2(S;\eta^0)\|^2\big|\{S_i\}_{i\in I_k^c}\right]\\
&\le \sup\nolimits_{\eta\in\mathcal{T}}\mathbb{E}_P\left[\|\psi_2(S;\eta)-\psi_2(S;\eta^0)\|^2\right]\\
&\lesssim r_N^2
\end{aligned}$$

due to arguments that are analogous to (25)–(29) presented in the proof of Lemma I.16. Because the event $\mathcal{E}_N$ holds with probability $1-\Delta_N = 1-o(1)$, we infer

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_2(S_i;\hat{\eta}^{I_k^c}) - \psi_2(S_i;\eta^0)\|^2 = O_P(r_N^2)$$

by Lemma I.12. Therefore, we have $\mathcal{I}_{k,A} = O_P(N^{-\frac{1}{2}} + r_N)$ by (43). In total, we thus have

$$\mathcal{I}_k = O_P\Big(N^{\max\left\{\frac{4}{p}-1,-\frac{1}{2}\right\}}\Big) + O_P\big(N^{-\frac{1}{2}} + r_N\big) = O_P\Big(N^{\max\left\{\frac{4}{p}-1,-\frac{1}{2}\right\}} + r_N\Big). \quad \square$$

Theorem I.21. *Suppose Assumption I.5 holds. Introduce the matrix*

$$\begin{aligned}&\hat{J}_{k,0}\\ &:= \Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widehat{R}^{I_k}_{X,i}(\widehat{R}^{I_k}_{A,i})^T\Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widehat{R}^{I_k}_{A,i}(\widehat{R}^{I_k}_{A,i})^T\Big)^{-1}\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widehat{R}^{I_k}_{A,i}(\widehat{R}^{I_k}_{X,i})^T\Big)^{-1}\\ &\quad\cdot\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widehat{R}^{I_k}_{X,i}(\widehat{R}^{I_k}_{A,i})^T\Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widehat{R}^{I_k}_{A,i}(\widehat{R}^{I_k}_{A,i})^T\Big)^{-1}.\end{aligned}$$

Let its average over $k\in[K]$ be

$$\hat{J}_0 := \frac{1}{K}\sum_{k=1}^K \hat{J}_{k,0}.$$

Define further the estimator

$$\hat{\sigma}^2 := \hat{J}_0\Big(\frac{1}{K}\sum_{k=1}^K\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{\beta},\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{\beta},\hat{\eta}^{I_k^c})\Big)\hat{J}_0^T$$

of σ^2 from Theorem 3.1, where $\hat{\beta}\in\{\hat{\beta}^{DML1},\hat{\beta}^{DML2}\}$. We then have $\hat{\sigma}^2 = \sigma^2 + O_P(\tilde{\rho}_N)$, where $\tilde{\rho}_N = N^{\max\left\{\frac{4}{p}-1,-\frac{1}{2}\right\}} + r_N$ is as in Definition I.4.

Proof of Theorem I.21. We derived $\hat{J}_{k,0} = J_0 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ in the proof of Theorem 3.1. Thus, $\hat{J}_0 = J_0 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ holds because K is a fixed number independent of N. To conclude the proof, it suffices to verify

$$\left\|\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{\beta},\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{\beta},\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\psi(S;\beta_0,\eta^0)\psi^T(S;\beta_0,\eta^0)\right]\right\| = O_P(\tilde{\rho}_N).$$

But this statement holds by Lemma I.20 because the dimensions of A and X are fixed. $\square$

Appendix J: Proofs of Section 4

Definition J.1. *Let $\gamma\geq 0$, and recall the scalar $\rho_N = r_N + N^{\frac{1}{2}}\lambda_N$ in Definition I.4. Introduce the function*

$$\begin{aligned}\overline{\psi}'(\cdot;b^\gamma,\eta^0) := &\ \widetilde{\psi}(\cdot;b^\gamma,\eta^0) + (\gamma-1)D_3\psi(\cdot;b^\gamma,\eta^0)\\ &+(\gamma-1)\big(\psi_1(\cdot;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5\\ &-(\gamma-1)D_3\big(\psi_2(\cdot;\eta^0) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5.\end{aligned}$$

Let

$$D_4 := \mathbb{E}_P\left[\overline{\psi}'(S;b^\gamma,\eta^0)(\overline{\psi}'(S;b^\gamma,\eta^0))^T\right],$$

and let the variance

$$\sigma^2(\gamma) := \big(D_1+(\gamma-1)D_2\big)^{-1}D_4\big(D_1^T+(\gamma-1)D_2^T\big)^{-1}.$$

Moreover, define the influence function

$$\overline{\psi}(\cdot;b^\gamma,\eta^0) := \sigma^{-1}(\gamma)\big(D_1+(\gamma-1)D_2\big)^{-1}\overline{\psi}'(\cdot;b^\gamma,\eta^0).$$

Proof of Theorem 4.1. This proof is based on Chernozhukov et al. [31]. The matrices $D_1+(\gamma-1)D_2$ and D_4 are invertible by Assumption I.5.4. Hence, $\sigma^2(\gamma)$ is invertible.

Subsequently, we show the stronger statement

$$\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^\gamma-b^\gamma)=\frac{1}{\sqrt{N}}\sum_{i=1}^N\overline{\psi}(S_i;b^\gamma,\eta^0)+O_P(\rho_N)\xrightarrow{d}\mathcal{N}(0,\mathbb{1}_{d\times d})\quad(N\to\infty), \tag{56}$$

where $\hat{b}^\gamma$ denotes the DML2 estimator $\hat{b}^{\gamma,\mathrm{DML2}}$ or its DML1 variant $\hat{b}^{\gamma,\mathrm{DML1}}$, and where $\overline{\psi}$ is as in Definition J.1. We first consider $\hat{b}^{\gamma,\mathrm{DML2}}$ and afterwards $\hat{b}^{\gamma,\mathrm{DML1}}$. Fix a sequence $\{P_N\}_{N\geq 1}$ such that $P_N\in\mathcal{P}_N$ for all $N\geq 1$. Because this sequence is chosen arbitrarily, it suffices to show

$$\begin{aligned}&\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^{\gamma,\mathrm{DML2}}-b^\gamma)\\&=\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N\overline{\psi}(S_i;b^\gamma,\eta^0)+O_{P_N}(\rho_N)\\&\xrightarrow{d}\mathcal{N}(0,\mathbb{1}_{d\times d})\quad(N\to\infty).\end{aligned}$$

We have

$$\begin{aligned}
&\hat{b}^{\gamma,\mathrm{DML2}}\\
&=\Big(\tfrac{1}{K}\textstyle\sum_{k=1}^K\big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T\big(\mathbb{1}+(\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\Big)^{-1}\\
&\quad\cdot\tfrac{1}{K}\textstyle\sum_{k=1}^K\big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T\big(\mathbb{1}+(\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}\\
&=\Bigg(\tfrac{1}{K}\textstyle\sum_{k=1}^K\Big(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad+(\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)\Bigg)^{-1}\\
&\quad\cdot\tfrac{1}{K}\textstyle\sum_{k=1}^K\Big(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad+(\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)
\end{aligned}\tag{57}$$

by (14). By Lemma I.17, we have

$$
\begin{aligned}
&\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&=\mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big),\\
&\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&=\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big),\\
&\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&=\mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big((X-m_X^0(W)\big)^T\Big]+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big).
\end{aligned}
$$

By Weyl's inequality and Slutsky's theorem, we hence have

$$
\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma,\mathrm{DML2}}-b^\gamma)\\
&=\Big(\big(D_1+(\gamma-1)D_2\big)^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\quad\cdot\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\tfrac{1}{\sqrt{n}}\Big(\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad\cdot\Big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})-\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)b^\gamma\Big)\\
&\qquad+(\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad\cdot\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad\cdot\Big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})-\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)b^\gamma\Big)\Big)\\
&=\Big(\big(D_1+(\gamma-1)D_2\big)^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\quad\cdot\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\Big(\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\hat{\eta}^{I_k^c})\\
&\quad+(\gamma-1)\cdot\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\cdot\big(\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\big)^{-1}\\
&\qquad\cdot\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\Big)
\end{aligned}
\tag{58}
$$

due to (57) because K and γ are constants independent of N and because $N=nK$ holds. Let $k\in[K]$. Next, we analyze the individual factors of the last summand in (58). By Lemma I.16, we have

$$
\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\\
&=\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)+\Big(\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})-\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)\Big)\\
&=\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)+O_{P_N}(\rho_N),
\end{aligned}
\tag{59}
$$

and

$$
\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\hat{\eta}^{I_k^c})\\
&=\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\eta^0)+\Big(\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\hat{\eta}^{I_k^c})-\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\eta^0)\Big)\\
&=\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;b^\gamma,\eta^0)+O_{P_N}(\rho_N),
\end{aligned}
\tag{60}
$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\\
&=\frac{1}{n}\sum_{i\in I_k}(\psi_1(S_i;\hat{\eta}^{I_k^c})-\psi_1(S_i;\eta^0))+\frac{1}{n}\sum_{i\in I_k}(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)])\\
&\quad+\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\\
&=O_{P_N}(N^{-\frac{1}{2}}\rho_N)+\frac{1}{n}\sum_{i\in I_k}(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)])+\mathbb{E}_{P_N}[\psi_1(S;\eta^0)].
\end{aligned}\tag{61}$$

We apply a series expansion to obtain

$$\begin{aligned}
&\Big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1}\\
&=\bigg(\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]+\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\psi_2(S_i;\eta^0)\big)\\
&\qquad+\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\bigg)^{-1}\\
&=\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\\
&\quad-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\psi_2(S_i;\eta^0)\big)\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\\
&\quad-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\qquad\cdot\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\\
&\quad+O_{P_N}\bigg(\Big\|\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\psi_2(S_i;\eta^0)\big)\Big\|^2\\
&\qquad\qquad+\Big\|\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\Big\|^2\bigg)\\
&=\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}\rho_N\big)+O_{P_N}\Big(O_{P_N}\big(N^{-1}\rho_N^2\big)+O_{P_N}(N^{-1})\Big)\\
&\quad-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\qquad\cdot\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\\
&=\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}\rho_N\big)\\
&\quad-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\qquad\cdot\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}
\end{aligned}\tag{62}$$

due to Lemma I.16, the Lindeberg–Feller CLT, the Cramer–Wold device, because $\rho_N\lesssim\delta_N^{\frac{1}{4}}$ holds by Lemma I.16, and because $\delta_N^{\frac{1}{4}}\geq N^{-\frac{1}{2}}$ holds by Assumption I.5. Thus, the last summand in (58) can be expressed as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\cdot\big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\big)^{-1}\cdot\frac{1}{n}\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\\
&=\sqrt{n}\Big(O_{P_N}\big(N^{-\frac{1}{2}}\rho_N\big)\\
&\qquad+\frac{1}{n}\sum_{i\in I_k}\big(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)+\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\Big)\\
&\quad\cdot\Big(\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}\rho_N\big)\\
&\qquad-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\qquad\cdot\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\Big)\Big(\frac{1}{n}\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)+O_{P_N}\big(N^{-\frac{1}{2}}\rho_N\big)\Big)\\
&=\frac{1}{\sqrt{n}}\sum_{i\in I_k}\big(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\mathbb{E}_{P_N}[\psi(S;b^\gamma,\eta^0)]\\
&\quad+\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)\\
&\quad-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\quad\cdot\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\mathbb{E}_{P_N}[\psi(S;b^\gamma,\eta^0)]+O_{P_N}(\rho_N)
\end{aligned}\tag{63}$$

due to (59)–(62), the Lindeberg–Feller CLT, and the Cramer–Wold device.

We combine (58) and (63) and obtain

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma,\mathrm{DML2}} - b^\gamma) \\
&= \Big(\big(D_1 + (\gamma-1)D_2\big)^{-1} + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K \tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\Big(\widetilde{\psi}(S_i; b^\gamma, \eta^0) + (\gamma-1)D_3\psi(S_i; b^\gamma, \eta^0) \\
&\qquad +(\gamma-1)\big(\psi_1(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)D_5 \\
&\qquad -(\gamma-1)D_3\big(\psi_2(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)D_5\Big) + O_{P_N}(\rho_N) \qquad (64) \\
&= \Big(\big(D_1 + (\gamma-1)D_2\big)^{-1}\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N \Big(\widetilde{\psi}(S_i; b^\gamma, \eta^0) + (\gamma-1)D_3\psi(S_i; b^\gamma, \eta^0) \\
&\qquad +(\gamma-1)\big(\psi_1(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)D_5 \\
&\qquad -(\gamma-1)D_3\big(\psi_2(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)D_5\Big) + O_{P_N}(\rho_N)
\end{aligned}$$

by the Lindeberg–Feller CLT and the Cramer–Wold device. We conclude our proof for the DML2 method by the Lindeberg–Feller CLT and the Cramer–Wold device.

Subsequently, we consider the DML1 method. It suffices to show that (56) holds uniformly over $P \in \mathcal{P}_N$. Fix a sequence $\{P_N\}_{N\geq 1}$ such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, it suffices to show

$$\begin{aligned}
&\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^{\gamma,\mathrm{DML1}} - b^\gamma) \\
&= \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N \overline{\psi}(S_i; b^\gamma, \eta^0) + O_{P_N}(\rho_N) \\
&\xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d\times d}) \quad (N\to\infty).
\end{aligned}$$

We have

$$\begin{aligned}
&\hat{b}^{\gamma,\mathrm{DML1}} \\
&= \tfrac{1}{K}\textstyle\sum_{k=1}^K \Big((\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T\big(\mathbb{1} + (\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\Big)^{-1} \\
&\qquad \cdot(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T\big(\mathbb{1} + (\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \\
&= \tfrac{1}{K}\textstyle\sum_{k=1}^K \Bigg(\Big(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad +(\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1} \qquad (65) \\
&\qquad \cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)\Bigg)^{-1} \\
&\quad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad +(\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1} \\
&\qquad \cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)
\end{aligned}$$

by (20). By Slutsky's theorem and Equation (65), we have

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma,\mathrm{DML1}} - b^\gamma)\\
&= \Big(\big(D_1 + (\gamma-1)D_2\big)^{-1} + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\quad \cdot \tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K \tfrac{1}{\sqrt{n}}\Big(\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad\qquad \cdot\Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T b^\gamma\Big)\\
&\qquad + (\gamma-1)\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad \cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad \cdot\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad \cdot\Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T b^\gamma\Big)\Big)\\
&= \Big(\big(D_1 + (\gamma-1)D_2\big)^{-1} + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\quad \cdot \tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K \sqrt{n}\Big(\tfrac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \hat{\eta}^{I_k^c})\\
&\quad + (\gamma-1)\cdot\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\cdot\big(\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\big)^{-1}\\
&\qquad \cdot\tfrac{1}{n}\textstyle\sum_{i\in I_k^c}\psi(S_i; b^\gamma, \hat{\eta}^{I_k^c})\Big).
\end{aligned}$$

The last expression above coincides with (58). Consequently, the same asymptotic analysis conducted for $\hat{b}^{\gamma,\mathrm{DML2}}$ can also be employed in this case. □

Lemma J.2. *Let $\gamma \geq 0$ and let $\varphi \in \{\psi, \widetilde{\psi}\}$. We have*

$$\frac{1}{n}\sum_{i\in I_k}\varphi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) = \mathbb{E}_P[\varphi(S;b^\gamma,\eta^0)] + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big).$$

Proof. We consider the case $\varphi = \psi$. We decompose

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\\
&= \tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\big)\\
&\quad + \tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c}) - \psi(S_i;b^\gamma,\eta^0)\big)\\
&\quad + \tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\eta^0) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\big).
\end{aligned} \tag{66}$$

Subsequently, we analyze the three terms in the above decomposition (66) individually. We have

$$\begin{aligned}
&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \tfrac{1}{n}\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\big\|\\
&\leq \big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}(A_i - \hat{m}_A^{I_k^c}(W_i))(X_i - \hat{m}_X^{I_k^c}(W_i))^T\big\|\|\hat{b}^\gamma - b^\gamma\|\\
&= \big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\big\|\|\hat{b}^\gamma - b^\gamma\|\\
&= \big\|\mathbb{E}_P[\psi_1(S;\eta^0)] + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\big\|\|\hat{b}^\gamma - b^\gamma\|
\end{aligned}$$

by Lemma I.17. Because $\|\hat{b}^\gamma - b^\gamma\| = O_P(N^{-\frac{1}{2}}\rho_N)$ holds by Theorem 4.1, we infer

$$\Big\|\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \frac{1}{n}\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\Big\| = O_P\big(N^{-\frac{1}{2}}\rho_N\big). \tag{67}$$

Due to (59) that was established in the proof of Theorem 4.1, we have

$$\frac{1}{n}\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\hat\eta^{I_k^c})-\psi(S_i;b^\gamma,\eta^0)\big)=O_P\big(N^{-\frac{1}{2}}\rho_N\big). \tag{68}$$

Due to the Lindeberg–Feller CLT and the Cramer–Wold device, we have

$$\frac{1}{n}\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\eta^0)-\mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\big)=O_P(N^{-\frac{1}{2}}). \tag{69}$$

We combine (66)–(69) to infer the claim for $\varphi=\psi$. The case $\varphi=\widetilde\psi$ can be analyzed analogously. ☐

Theorem J.3. *Suppose Assumption I.5 holds. Recall the score functions introduced in Definition I.1, and let* $\hat b^\gamma\in\{\hat b^{\gamma,DML1},\hat b^{\gamma,DML2}\}$. *Introduce the matrices*

$$\begin{aligned}
\hat D_1^k &:= \tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_3(S_i;\hat\eta^{I_k^c}),\\
\hat D_2^k &:= \tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S;\hat\eta^{I_k^c})\Big(\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat\eta^{I_k^c})\Big)^{-1}\tfrac{1}{n}\sum_{i\in I_k}\psi_1^T(S_i;\hat\eta^{I_k^c}),\\
\hat D_3^k &:= \tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat\eta^{I_k^c})\Big(\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat\eta^{I_k^c})\Big)^{-1},\\
\hat D_5^k &:= \Big(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_2(S_i;\hat\eta^{I_k^c})\Big)^{-1}\tfrac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat b^\gamma,\hat\eta^{I_k^c}).
\end{aligned}$$

Let furthermore

$$\begin{aligned}
\widehat{\widetilde\psi}'(\cdot;\hat b^\gamma,\hat\eta^{I_k^c}) := \widetilde\psi(\cdot;\hat b^\gamma,\hat\eta^{I_k^c}) &+(\gamma-1)\hat D_3^k\psi(\cdot;\hat b^\gamma,\hat\eta^{I_k^c})\\
&+(\gamma-1)\big(\psi_1(\cdot;\hat\eta^{I_k^c})-\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat\eta^{I_k^c})\big)\hat D_5^k\\
&-(\gamma-1)\hat D_3^k\big(\psi_2(\cdot;\hat\eta^{I_k^c})-\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_2(S_i;\hat\eta^{I_k^c})\big)\hat D_5^k
\end{aligned}$$

and

$$\hat D_4^k := \frac{1}{n}\sum_{i\in I_k}\widehat{\widetilde\psi}'(S_i;\hat b^\gamma,\hat\eta^{I_k^c})\big(\widehat{\widetilde\psi}'(S_i;\hat b^\gamma,\hat\eta^{I_k^c})\big)^T.$$

Define the estimators

$$\hat D_1 := \frac{1}{K}\sum_{k=1}^K\hat D_1^k,\quad \hat D_2 := \frac{1}{K}\sum_{k=1}^K\hat D_2^k,\quad \textit{and}\quad \hat D_4 := \frac{1}{K}\sum_{k=1}^K\hat D_4^k.$$

We estimate the asymptotic variance covariance matrix $\sigma^2(\gamma)$ *in Theorem 4.1 by*

$$\hat\sigma^2(\gamma) := \big(\hat D_1+(\gamma-1)\hat D_2\big)^{-1}\hat D_4\big(\hat D_1^T+(\gamma-1)\hat D_2^T\big)^{-1}.$$

Then we have $\hat\sigma^2(\gamma)=\sigma^2(\gamma)+O_P\big(\tilde\rho_N+N^{-\frac{1}{2}}(1+\rho_N)\big)$, *where we have* $\tilde\rho_N = N^{\max\left\{\frac{4}{p}-1,-\frac{1}{2}\right\}}+r_N$ *is as in Definition I.4.*

Proof of Theorem J.3. This proof is based on Chernozhukov et al. [31]. We already verified

$$\hat{D}_1 = D_1 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \quad \text{and} \quad \hat{D}_2 = D_2 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$$

in the proof of Theorem 4.1 because K is a fixed integer independent of N. Thus, we have

$$\big(\hat{D}_1 + (\gamma-1)\hat{D}_2\big)^{-1} = \big(D_1 + (\gamma-1)D_2\big)^{-1} + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$$

by Weyl's inequality. Moreover, we have $\hat{D}_3^k = D_3 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ by Lemma I.17.

Subsequently, we argue that $\hat{D}_5^k = D_5 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ holds. Due to Lemma I.17 and Weyl's inequality, we have

$$\frac{1}{n}\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c}) = \mathbb{E}_P[\psi_1(S;\eta^0)] + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$$

and

$$\Big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1} = \mathbb{E}_P[\psi_2(S;\eta^0)]^{-1} + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big). \tag{70}$$

Due to (70), it suffices to show

$$\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) = \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)] + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \tag{71}$$

to infer $\hat{D}_5^k = D_5 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$. But (71) holds due to Lemma J.2. To conclude the theorem, it remains verify $\hat{D}_4^k = D_4 + O_P(\tilde{\rho}_N)$. We have

$$\begin{aligned}
&\|\hat{D}_4^k - D_4\| \\
\le &\Big\|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\widetilde{\psi}(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\Big\| \\
&+ (\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T \\
&\qquad - \mathbb{E}_P\big[\widetilde{\psi}(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\big]D_3^T\Big\| \\
&+ (\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\qquad - D_3\,\mathbb{E}_P\big[\psi(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\Big\| \\
&+ (\gamma-1)^2\Big\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T
\end{aligned}$$

$$
\begin{aligned}
&- D_3 \, \mathbb{E}_P \left[\psi(S; b^\gamma, \eta^0) \psi^T(S; b^\gamma, \eta^0) \right] D_3^T \Bigg\| \\
&+ (\gamma - 1) \Bigg\| \frac{1}{n} \sum_{i \in I_k} \widetilde{\psi}(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) D_5^T \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \\
&\quad - \mathbb{E}_P \left[\widetilde{\psi}(S; b^\gamma, \eta^0) D_5^T \big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \right] \Bigg\| \\
&+ (\gamma - 1) \Bigg\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \widetilde{\psi}^T(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) \\
&\quad - \mathbb{E}_P \left[\big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \widetilde{\psi}^T(S; b^\gamma, \eta^0) \right] \Bigg\| \\
&+ (\gamma - 1)^2 \Bigg\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \\
&\qquad \cdot D_5^T \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \\
&\quad - \mathbb{E}_P \left[\big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 D_5^T \big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \right] \Bigg\| \\
&+ (\gamma - 1) \Bigg\| \frac{1}{n} \sum_{i \in I_k} D_3 \big(\psi_2(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S; \eta^0)] \big) D_5 \widetilde{\psi}^T(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) \\
&\quad - D_3 \, \mathbb{E}_P \left[\big(\psi_2(S; \eta^0) - \mathbb{E}_P[\psi_2(S; \eta^0)] \big) D_5 \widetilde{\psi}^T(S; b^\gamma, \eta^0) \right] \Bigg\| \\
&+ (\gamma - 1) \Bigg\| \frac{1}{n} \sum_{i \in I_k} \widetilde{\psi}(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) D_5^T \big(\psi_2(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S; \eta^0)] \big)^T D_3^T \\
&\quad - \mathbb{E}_P \left[\widetilde{\psi}(S; b^\gamma, \eta^0) D_5^T \big(\psi_2(S; \eta^0) - \mathbb{E}_P[\psi_2(S; \eta^0)] \big)^T \right] D_3^T \Bigg\| \\
&+ (\gamma - 1)^2 \Bigg\| \frac{1}{n} \sum_{i \in I_k} D_3 \psi(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) D_5^T \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \\
&\quad - D_3 \, \mathbb{E}_P \left[\psi(S; b^\gamma, \eta^0) D_5^T \big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big)^T \right] \Bigg\| \\
&+ (\gamma - 1)^2 \Bigg\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \psi^T(S_i; \hat{b}^\gamma, \hat{\eta}^{I_k^c}) D_3^T \\
&\quad - \mathbb{E}_P \left[\big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \psi^T(S; b^\gamma, \eta^0) \right] D_3^T \Bigg\| \\
&+ (\gamma - 1)^2 \Bigg\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_1(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5 \\
&\qquad \cdot D_5^T \big(\psi_2(S_i; \hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S; \eta^0)] \big)^T D_3^T \\
&\quad - \mathbb{E}_P \Big[\big(\psi_1(S; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)] \big) D_5
\end{aligned}
$$

$$
\begin{aligned}
&\cdot D_5^T\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\bigg\| \\
&+(\gamma-1)^2\bigg\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T D_3^T \\
&\quad -D_3\,\mathbb{E}_P\Big[\psi(S_i;b^\gamma,\eta^0)D_5^T\big(\psi_2(S_i;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\bigg\| \\
&+(\gamma-1)^2\bigg\|\frac{1}{n}\sum_{i\in I_k}D_3\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T \\
&\quad -D_3\,\mathbb{E}_P\big[\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5\psi^T(S;b^\gamma,\eta^0)\big]D_3^T\bigg\| \\
&+(\gamma-1)^2\bigg\|\frac{1}{n}\sum_{i\in I_k}D_3\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad\cdot D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T \\
&\quad -D_3\,\mathbb{E}_P\Big[\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad\cdot D_5^T\big(\psi_1(S;\eta^0)-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\Big]\bigg\| \\
&+(\gamma-1)^2\bigg\|\frac{1}{n}\sum_{i\in I_k}D_3\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad\cdot D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T D_3^T \\
&\quad -D_3\,\mathbb{E}_P\Big[\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad\cdot D_5^T\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\bigg\| \\
&+O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \\
=:&\sum_{i=1}^{16}\mathcal{I}_i+O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}
$$

by the triangle inequality and the results derived so far. Subsequently, we bound the terms $\mathcal{I}_1,\dots,\mathcal{I}_{16}$ individually. Because all these terms consist of norms of matrices of fixed size, it suffices to bound the individual matrix entries. Let j,l,t,r be natural numbers not exceeding the dimensions of the respective object they index. By Lemma I.20, we have

$$
\bigg|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}_l(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})-\mathbb{E}_P\big[\widetilde{\psi}_j(S;b^\gamma,\eta^0)\widetilde{\psi}_l(S;b^\gamma,\eta^0)\big]\bigg|=O_P(\tilde{\rho}_N),
$$

which implies $\mathcal{I}_1=O_P(\tilde{\rho}_N)$. By Lemma I.20, we have

$$
\bigg|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi_l(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})-\mathbb{E}_P\big[\widetilde{\psi}_j(S;b^\gamma\eta^0)\psi_l(S;\beta_0,\eta^0)\big]\bigg|=O_P(\tilde{\rho}_N),
$$

which implies $\mathcal{I}_2 = O_P(\tilde{\rho}_N) = \mathcal{I}_3$ due to

$$\begin{aligned}&\big\|\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]D_3^T\big\| \\ \leq &\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]\big\|\|D_3\|\end{aligned}$$

and

$$\begin{aligned}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\ &\quad -D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\right]\big\| \\ \leq &\|D_3\|\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\right]\Big\|.\end{aligned}$$

By Lemma I.20, we have

$$\left|\frac{1}{n}\sum_{i\in I_k}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi_l(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_j(S;\beta_0,\eta^0)\psi_l(S;\beta_0,\eta^0)]\right| = O_P(\tilde{\rho}_N),$$

which implies $\mathcal{I}_4 = O_P(\tilde{\rho}_N)$ due to

$$\begin{aligned}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T \\ &\quad -D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]D_3^T\big\| \\ \leq &\|D_3\|^2\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]\big\|.\end{aligned}$$

By Lemma I.20, we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{l,t} - \mathbb{E}_P\left[\widetilde{\psi}_j(S;b^\gamma,\eta^0)\big(\psi_1(S;\eta^0)\big)_{l,t}\right]\big| \\ = &O_P(\tilde{\rho}_N),\end{aligned}$$

which implies $\mathcal{I}_5 = O_P(\tilde{\rho}_N)$ because we have

$$\begin{aligned}&\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T \\ &\quad -\mathbb{E}_P\Big[\widetilde{\psi}(S;b^\gamma,\eta^0)D_5^T\big(\psi_1(S;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\Big]\Big\| \\ \leq &\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)\right]\big\| \\ &\quad +\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)\right]\big\|\|D_5\|\|\mathbb{E}_P[\psi_1(S;\eta^0)]\|,\end{aligned}$$

where the last summand is $O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ by Lemma J.2, and we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{j,l} \\ &\quad -\big(\mathbb{E}_P[\widetilde{\psi}(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)]\big)_{j,l}\big| \\ = &\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,l}\widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\ &\quad -D_5^T\,\mathbb{E}_P\left[\big(\psi_1(S;\eta^0)\big)_{\cdot,l}\widetilde{\psi}_j(S;b^\gamma,\eta^0)\right]\big| \\ \leq &\|D_5\|\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,l}\widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\ &\quad -\mathbb{E}_P\left[\big(\psi_1(S;\eta^0)\big)_{\cdot,l}\widetilde{\psi}_j(S;b^\gamma,\eta^0)\right]\big\|.\end{aligned}$$

The term $\mathcal{I}_6$ can be bounded analogously to $\mathcal{I}_5$. By Lemma I.20, we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,l}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{t,r} - \mathbb{E}_P\left[\big(\psi_1(S;\eta^0)\big)_{j,l}\big(\psi_1(S;\eta^0)\big)_{l,t}\right]\big| \\ = &O_P(\tilde{\rho}_N),\end{aligned}$$

which implies $\mathcal{I}_7 = O_P(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}
&\Big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big) D_5 D_5^T \big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T \\
&\quad - \mathbb{E}_P\Big[\big(\psi_1(S;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big) D_5 D_5^T \big(\psi_1(S;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\Big]\Big\| \\
\le\ & \big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \psi_1(S_i;\hat{\eta}^{I_k^c}) D_5 D_5^T \psi_1^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\psi_1(S;\eta^0) D_5 D_5^T \psi_1^T(S;\eta^0)\big]\big\| \\
&\quad +2\big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big\| \|D_5\|^2 \|\mathbb{E}_P[\psi_1(S;\eta^0)]\| \\
=\ & \big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \psi_1(S_i;\hat{\eta}^{I_k^c}) D_5 D_5^T \psi_1^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\psi_1(S;\eta^0) D_5 D_5^T \psi_1^T(S;\eta^0)\big]\big\| \\
&\quad +O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}$$

by Lemma I.17, and we have

$$\begin{aligned}
&\Big| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \big(\psi_1(S_i;\hat{\eta}^{I_k^c}) D_5 D_5^T \psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{j,r} \\
&\quad -\big(\mathbb{E}_P\big[\psi_1(S;\eta^0) D_5 D_5^T \psi_1^T(S;\eta^0)\big]\big)_{j,r}\Big| \\
=\ & \Big| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot} D_5 D_5^T (\psi_1^T(S_i;\hat{\eta}^{I_k^c}))_{\cdot,r} \\
&\quad - \mathbb{E}_P\Big[\big(\psi_1(S;\eta^0)\big)_{j,\cdot} D_5 D_5^T \big(\psi_1^T(S;\eta^0)\big)_{\cdot,r}\Big]\Big| \\
=\ & \big| \tfrac{1}{n} \textstyle\sum_{i\in I_k} D_5^T \big(\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r} \big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot} D_5 \\
&\quad - \mathbb{E}_P\big[D_5^T \big(\psi_1^T(S;\eta^0)\big)_{\cdot,r} \big(\psi_1(S;\eta^0)\big)_{j,\cdot} D_5\big]\big| \\
\le\ & \Big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \big(\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r} \big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot} - \mathbb{E}_P\Big[\big(\psi_1^T(S;\eta^0)\big)_{\cdot,r} \big(\psi_1(S;\eta^0)\big)_{j,\cdot}\Big]\Big\| \\
&\quad \cdot\|D_5\|^2.
\end{aligned}$$

Next, we bound $\mathcal{I}_8$. By Lemma I.20, we have

$$\begin{aligned}
&\big| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \widetilde{\psi}_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{l,t} - \mathbb{E}_P\big[\widetilde{\psi}_j(S_i;b^\gamma,\eta^0)\big(\psi_2(S;\eta^0)\big)_{l,t}\big]\big| \\
&= O_P(\tilde{\rho}_N),
\end{aligned}$$

which implies $\mathcal{I}_8 = O_{P_N}(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}
&\big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} D_3 \big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big) D_5 \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\quad -D_3\,\mathbb{E}_P\big[\big(\psi_2(S;\eta^0) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big) D_5 \widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\big\| \\
\le\ & \Big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} D_3 \psi_2(S_i;\hat{\eta}^{I_k^c}) D_5 \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\qquad -D_3\,\mathbb{E}_P\big[\psi_2(S;\eta^0) D_5 \widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\Big\| \\
&\quad +\big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} D_3\,\mathbb{E}_P[\psi_2(S;\eta^0)] D_5 \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\qquad -D_3\,\mathbb{E}_P[\psi_2(S;\eta^0)] D_5\,\mathbb{E}_P\big[\widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\big\| \\
\le\ & \|D_3\| \big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \psi_2(S_i;\hat{\eta}^{I_k^c}) D_5 \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\qquad - \mathbb{E}_P\big[\psi_2(S;\eta^0) D_5 \widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\big\| \\
&\quad +\|D_3\| \|\mathbb{E}_P[\psi_2(S;\eta^0)]\| \|D_5\| \\
&\quad \cdot\big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\big\| \\
\le\ & \|D_3\| \big\| \tfrac{1}{n} \textstyle\sum_{i\in I_k} \psi_2(S_i;\hat{\eta}^{I_k^c}) D_5 \widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\quad - \mathbb{E}_P\big[\psi_2(S;\eta^0) D_5 \widetilde{\psi}^T(S;b^\gamma,\eta^0)\big]\big\| + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}$$

by Lemma J.2, and we have

$$\begin{aligned}
&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})D_5\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big)_{j,t}\\
&\quad-\big(\mathbb{E}_P\left[\psi_2(S;\eta^0)D_5\widetilde{\psi}^T(S;b^\gamma,\eta^0)\right]\big)_{j,t}\big|\\
=&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}D_5\widetilde{\psi}_t(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\\
&\quad-\mathbb{E}_P\left[\big(\psi_2(S;\eta^0)\big)_{j,\cdot}D_5\widetilde{\psi}_t(S;b^\gamma,\eta^0)\right]\big|\\
=&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}_t(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}D_5\\
&\quad-\mathbb{E}_P\left[\widetilde{\psi}_t(S;b^\gamma,\eta^0)\big(\psi_2(S;\eta^0)\big)_{j,\cdot}D_5\right]\big|\\
\le&\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\widetilde{\psi}_t(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}-\mathbb{E}_P\left[\widetilde{\psi}_t(S;b^\gamma,\eta^0)\big(\psi_2(S;\eta^0)\big)_{j,\cdot}\right]\Big\|\\
&\quad\cdot\|D_5\|.
\end{aligned}$$

The term $\mathcal{I}_9$ can be bounded analogously to $\mathcal{I}_8$. Next, we bound $\mathcal{I}_{10}$. By Lemma I.20, we have

$$\begin{aligned}
&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{l,t}-\mathbb{E}_P\left[\psi_j(S;b^\gamma,\eta^0)\big(\psi_1(S;\eta^0)\big)_{l,t}\right]\big|\\
&=O_P(\tilde{\rho}_N),
\end{aligned}$$

which implies $\mathcal{I}_{10}=O_{P_N}(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}
&\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\\
&\quad-D_3\,\mathbb{E}_P\Big[\psi(S;b^\gamma,\eta^0)D_5^T\big(\psi_1(S;\eta^0)-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\Big]\Big\|\\
\le&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c})\\
&\qquad-D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)\right]\big\|\\
&\quad+\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\,\mathbb{E}_{P_N}[\psi_1^T(S;\eta^0)]\\
&\qquad-D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)D_5^T\,\mathbb{E}_P[\psi_1^T(S;\eta^0)]\right]\big\|\\
\le&\|D_3\|\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)\right]\big\|\\
&\quad+\|D_3\|\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi(S;b^\gamma,\eta^0)\big\|\|D_5\|\|\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\|\\
\le&\|D_3\|\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)\right]\big\|\\
&\quad+O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}$$

by Lemma J.2, and we have

$$\begin{aligned}
&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{j,t}\\
&\quad-\big(\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)D_5^T\psi_1^T(S;\eta^0)\right]\big)_{j,t}\big|\\
=&\Big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,t}\\
&\quad-\mathbb{E}_P\Big[\psi_j(S;b^\gamma,\eta^0)D_5^T\big(\psi_1^T(S;\eta^0)\big)_{\cdot,t}\Big]\Big|\\
=&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_5^T\big(\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,t}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\\
&\quad-\mathbb{E}_P\left[D_5^T\big(\psi_1^T(S;\eta^0)\big)_{\cdot,t}\psi_j(S;b^\gamma,\eta^0)\right]\big|\\
\le&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,t}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})-\mathbb{E}_P\left[\big(\psi_1^T(S;\eta^0)\big)_{\cdot,t}\psi_j(S;b^\gamma,\eta^0)\right]\big\|\\
&\quad\cdot\|D_5\|.
\end{aligned}$$

The term $\mathcal{I}_{11}$ can be bounded analogously to $\mathcal{I}_{10}$. Next, we bound $\mathcal{I}_{12}$. By Lemma I.20, we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c}))_{j,l}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{t,r}-\mathbb{E}_P\left[\big(\psi_1(S;\eta^0)\big)_{j,l}\big(\psi_2(S;\eta^0)\big)_{t,r}\right]\big|\\&=O_P(\tilde{\rho}_N),\end{aligned}$$

which implies $\mathcal{I}_{12}=O_{P_N}(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5\\&\quad\cdot D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_3^T\\&\quad-\mathbb{E}_P\left[\big(\psi_1(S;\eta^0)-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5D_5^T\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)\right]D_3^T\big\|\\ \le{}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})D_3^T\\&\qquad-\mathbb{E}_P\left[\psi_1(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\right]D_3^T\big\|\\&\quad+\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\,\mathbb{E}_P[\psi_2^T(S;\eta^0)]D_3^T\\&\qquad-\mathbb{E}_P\left[\psi_1(S;\eta^0)D_5D_5^T\,\mathbb{E}_P[\psi_2^T(S;\eta^0)]\right]D_3^T\big\|\\&\quad+\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\mathbb{E}_P[\psi_1(S;\eta^0)]D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})D_3^T\\&\qquad-\mathbb{E}_P\left[\mathbb{E}_P[\psi_1(S;\eta^0)]D_5D_5^T\psi_2^T(S;\eta^0)\right]D_3^T\big\|\\ \le{}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P\left[\psi_1(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\right]\big\|\\&\quad\cdot\|D_3\|\\&\quad+\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big\|\|D_5\|^2\|\mathbb{E}_P[\psi_2(S;\eta^0)]\|\|D_3\|\\&\quad+\|\mathbb{E}_P[\psi_1(S;\eta^0)]\|\|D_5\|^2\|D_3\|\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_2(S;\eta^0)]\big\|\\ \le{}&\big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)]\big\|\\&\quad\cdot\|D_3\|+O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\end{aligned}$$

by Lemma I.17, and we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{j,r}\\&\quad-\big(\mathbb{E}_P\left[\psi_1(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\right]\big)_{j,r}\big|\\ ={}&\Big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}D_5D_5^T\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r}\\&\quad-\mathbb{E}_P\Big[\big(\psi_1(S;\eta^0)\big)_{j,\cdot}D_5D_5^T\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\Big]\Big|\\ ={}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_5^T\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}D_5\\&\quad-\mathbb{E}_P\left[D_5^T\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\big(\psi_1(S;\eta^0)\big)_{j,\cdot}D_5\right]\big|\\ \le{}&\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})\big)_{j,\cdot}-\mathbb{E}_P\Big[\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\big(\psi_1(S;\eta^0)\big)_{j,\cdot}\Big]\Big\|\\&\quad\cdot\|D_5\|^2.\end{aligned}$$

Next, we bound $\mathcal{I}_{13}$. By Lemma I.20, we have

$$\begin{aligned}&\big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{t,r}-\mathbb{E}_P\left[\psi_j(S;b^\gamma,\eta^0)\big(\psi_2(S;\eta^0)\big)_{t,r}\right]\big|\\&=O_P(\tilde{\rho}_N),\end{aligned}$$

which implies $\mathcal{I}_{13} = O_P(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}
&\Big\| \tfrac{1}{n}\sum_{i\in I_k} D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T D_3^T \\
&\qquad -D_3\,\mathbb{E}_P\Big[\psi(S;b^\gamma,\eta^0)D_5^T\big(\psi_2(S;\eta^0)-\mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\Big\| \\
&\le \Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\psi(S;b^\gamma,\eta^0)D_5^T\psi_2^T(S;\eta^0)\big]\Big\| \\
&\quad \cdot\|D_3\|^2 \\
&\quad +\|D_3\|^2\|D_5\|\|\mathbb{E}_P[\psi_2(S;\eta^0)]\|\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\Big\| \\
&= \Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\psi(S;b^\gamma,\eta^0)D_5^T\psi_2^T(S;\eta^0)\big]\Big\| \\
&\quad \cdot\|D_3\|^2 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}$$

by Lemma J.2, and we have

$$\begin{aligned}
&\Big|\tfrac{1}{n}\sum_{i\in I_k}\big(\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{j,r} \\
&\qquad -\mathbb{E}_P\Big[\big(\psi(S;b^\gamma,\eta^0)D_5^T\psi_2^T(S;\eta^0)\big)_{j,r}\Big]\Big| \\
&= \Big|\tfrac{1}{n}\sum_{i\in I_k}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r} \\
&\qquad -\mathbb{E}_P\Big[\psi_j(S;b^\gamma,\eta^0)D_5^T\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\Big]\Big| \\
&= \big|\tfrac{1}{n}\sum_{i\in I_k}D_5^T(\psi_2^T(S_i;\hat{\eta}^{I_k^c}))_{\cdot,r}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\qquad -\mathbb{E}_P\big[D_5^T(\psi_2^T(S;\eta^0))_{\cdot,r}\psi_j(S;b^\gamma,\eta^0)\big]\big| \\
&\le \Big\|\tfrac{1}{n}\sum_{i\in I_k}\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{\cdot,r}\psi_j(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\Big[\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\psi_j(S;b^\gamma,\eta^0)\Big]\Big\| \\
&\quad \cdot\|D_5\|.
\end{aligned}$$

The term $\mathcal{I}_{14}$ can be bounded analogously to $\mathcal{I}_{13}$. The term $\mathcal{I}_{15}$ can be bounded analogously to $\mathcal{I}_{12}$. Last, we bound the term $\mathcal{I}_{16}$. By Lemma I.20, we have

$$\begin{aligned}
&\Big|\tfrac{1}{n}\sum_{i\in I_k}\big(\psi_2^T(S_i;\hat{\eta}^{I_k^c})\big)_{t,r}\big(\psi_2(S_i;\hat{\eta}^{I_k^c})\big)_{j,l} - \mathbb{E}_P\Big[\big(\psi_2^T(S;\eta^0)\big)_{t,r}\big(\psi_2(S;\eta^0)\big)_{j,l}\Big]\Big| \\
&= O_P(\tilde{\rho}_N),
\end{aligned}$$

which implies $\mathcal{I}_{16} = O_P(\tilde{\rho}_N)$. Indeed, we have

$$\begin{aligned}
&\Big\|\tfrac{1}{n}\sum_{i\in I_k}D_3\big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad \cdot D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T D_3^T \\
&\qquad -D_3\,\mathbb{E}_P\Big[\big(\psi_2(S;\eta^0) - \mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)D_5 \\
&\qquad \cdot D_5^T\big(\psi_2(S;\eta^0) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\Big\| \\
&\le \|D_3\|^2\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) \\
&\qquad -\mathbb{E}_P\big[\psi_2(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\big]\Big\| \\
&\quad +2\|D_3\|^2\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\,\mathbb{E}_{P_N}\big[\psi_2^T(S;\eta^0)\big] \\
&\qquad -\mathbb{E}_P\big[\psi_2(S;\eta^0)D_5D_5^T\,\mathbb{E}_P[\psi_2^T(S;\eta^0)]\big]\Big\| \\
&\le \|D_3\|^2\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) \\
&\qquad -\mathbb{E}_P\big[\psi_2(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\big]\Big\| \\
&\quad +2\|D_3\|^2\|D_5\|^2\|\mathbb{E}_P[\psi_2(S;\eta^0)]\|\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\Big\| \\
&= \|D_3\|^2\Big\|\tfrac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) \\
&\qquad -\mathbb{E}_P\big[\psi_2(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\big]\Big\| + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned}$$

by Lemma I.17, and we have

$$\begin{aligned}
&\Big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2(S_i;\hat\eta^{I_k^c})D_5D_5^T\psi_2^T(S_i;\hat\eta^{I_k^c})\big)_{j,r}\\
&\quad-\big(\mathbb{E}_P\left[\psi_2(S;\eta^0)D_5D_5^T\psi_2^T(S;\eta^0)\right]\big)_{j,r}\Big|\\
&=\Big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2(S_i;\hat\eta^{I_k^c})\big)_{j,\cdot}D_5D_5^T\big(\psi_2^T(S_i;\hat\eta^{I_k^c})\big)_{\cdot,r}\\
&\quad-\mathbb{E}_P\left[\big(\psi_2(S;\eta^0)\big)_{j,\cdot}D_5D_5^T\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}\right]\Big|\\
&=\Big|\tfrac{1}{n}\textstyle\sum_{i\in I_k}D_5^T\big(\psi_2^T(S_i;\hat\eta^{I_k^c})\big)_{\cdot,r}\big(\psi_2(S_i;\hat\eta^{I_k^c})\big)_{j,\cdot}D_5\\
&\quad-D_5^T\,\mathbb{E}_P\left[\big(\psi_2^T(S;\eta^0)\big)_{\cdot,r}(\psi_2(S;\eta^0))_{j,\cdot}\right]D_5\Big|\\
&\le\Big\|\tfrac{1}{n}\textstyle\sum_{i\in I_k}\big(\psi_2^T(S_i;\hat\eta^{I_k^c})\big)_{\cdot,r}(\psi_2(S_i;\hat\eta^{I_k^c}))_{j,\cdot}-\mathbb{E}_P\left[(\psi_2^T(S;\eta^0))_{\cdot,r}\big(\psi_2(S;\eta^0)\big)_{j,\cdot}\right]\Big\|\\
&\quad\cdot\|D_5\|^2.\quad\square
\end{aligned}$$

Proof of Proposition 4.2. The statement of Proposition 4.2 can be reformulated as

$$\sqrt{N}|b^{\gamma_N}-\beta_0|\to\begin{cases}0, & \text{if } \gamma_N=\Omega(\sqrt{N}) \text{ and } \gamma_N\notin\Theta(\sqrt{N})\\ C, & \text{if } \gamma_N=\Theta(\sqrt{N})\\ \infty, & \text{if } \gamma_N=o(\sqrt{N})\end{cases}$$

using the Bachmann–Landau notation, which is presented in Lattimore and Szepesvári [58], for instance.

Introduce the matrices

$$\begin{aligned}
F_1&:=\mathbb{E}_P[R_XR_Y],\\
F_2&:=\mathbb{E}_P\left[R_XR_X^T\right],\\
G_1&:=\mathbb{E}_P\left[R_XR_A^T\right]\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\mathbb{E}_P[R_AR_Y],\\
G_2&:=\mathbb{E}_P\left[R_XR_A^T\right]\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\mathbb{E}_P\left[R_AR_X^T\right].
\end{aligned}$$

We have

$$\sqrt{N}|b^{\gamma_N}-\beta_0|=\sqrt{N}\Big|\big(F_2+(\gamma_N-1)G_2\big)^{-1}\big(F_1+(\gamma_N-1)G_1\big)-G_2^{-1}G_1\Big|.$$

First, we assume that the sequence $\{\gamma_N\}_{N\ge1}$ diverges to $+\infty$ as $N\to\infty$, so that γ_N-1 is bounded away from 0 for N large enough. By Henderson and Searle [50, Section 3], we have

$$\begin{aligned}
&\big(F_2+(\gamma_N-1)G_2\big)^{-1}\\
&=\tfrac{1}{\gamma_N-1}G_2^{-1}-\Big(\mathbb{1}+\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\Big)^{-1}\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\tfrac{1}{\gamma_N-1}G_2^{-1}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\sqrt{N}|b^{\gamma_N}-\beta_0|\\
&=\tfrac{\sqrt{N}}{\gamma_N-1}\Big|G_2^{-1}F_1-\big(\mathbb{1}+\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\big)^{-1}\tfrac{1}{\gamma_N-1}G_2^{-1}F_2G_2^{-1}F_1\\
&\qquad-\big(\mathbb{1}+\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\big)^{-1}G_2^{-1}F_2G_2^{-1}G_1\Big|
\end{aligned}$$

and infer our claim because we have

$$\begin{aligned}
&G_2^{-1}F_1-\big(\mathbb{1}+\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\big)^{-1}\tfrac{1}{\gamma_N-1}G_2^{-1}F_2G_2^{-1}F_1\\
&\quad-\big(\mathbb{1}+\tfrac{1}{\gamma_N-1}G_2^{-1}F_2\big)^{-1}G_2^{-1}F_2G_2^{-1}G_1\\
&=O(1).
\end{aligned}$$

Next, we assume that the sequence $\{\gamma_N\}_{N\geq 1}$ is bounded. We have

$$|b^{\gamma_N} - \beta_0| = \left|\big(F_2 + (\gamma_N - 1)G_2\big)^{-1}\big(F_1 + (\gamma_N - 1)G_1\big) - G_2^{-1}G_1\right| = O(1),$$

which concludes the proof. ☐

Proof of Theorem 4.3. We show that

$$P\big(\hat{\sigma}^2(\gamma_N) + N(\hat{b}^{\gamma_N} - \hat{\beta})^2 \leq \hat{\sigma}^2\big) \leq P(|\Xi_N| \geq C_N)$$

holds for some random variable Ξ_N satisfying $\Xi_N = O_P(1)$ and for some sequence $\{C_N\}_{N\geq 1}$ of non-negative numbers diverging to $+\infty$ as $N \to \infty$.

For real numbers a and b, observe that we have

$$\sqrt{|a|^2 + |b|^2} \geq \frac{1}{2}|a| + \frac{1}{2}|b|$$

due to

$$\frac{3}{4}\Big(|a|^2 + |b|^2 - \frac{2}{3}|a||b|\Big) \geq \frac{3}{4}(|a| - |b|)^2 \geq 0.$$

Thus, we have

$$\begin{aligned} &P\big(\hat{\sigma}^2(\gamma_N) + N(\hat{b}^{\gamma_N} - \hat{\beta})^2 \leq \hat{\sigma}^2\big) \\ &= P\Big(\sqrt{\hat{\sigma}^2(\gamma_N) + N(\hat{b}^{\gamma_N} - \hat{\beta})^2} \leq \hat{\sigma}\Big) \\ &\leq P\big(\hat{\sigma}(\gamma_N) + \sqrt{N}|\hat{b}^{\gamma_N} - \hat{\beta}| \leq 2\hat{\sigma}\big). \end{aligned}$$

By the reverse triangle inequality, we have

$$\begin{aligned} &|\hat{b}^{\gamma_N} - \hat{\beta}| \\ &= |\hat{b}^{\gamma_N} - b^{\gamma_N} + b^{\gamma_N} - \beta_0 + \beta_0 - \hat{\beta}| \\ &\geq |b^{\gamma_N} - \beta_0| - |\hat{b}^{\gamma_N} - b^{\gamma_N}| - |\beta_0 - \hat{\beta}|. \end{aligned}$$

Thus, we have

$$\begin{aligned} &P\big(\hat{\sigma}^2(\gamma_N) + N(\hat{b}^{\gamma_N} - \hat{\beta})^2 \leq 2\hat{\sigma}^2\big) \\ &\leq P\big(\hat{\sigma}(\gamma_N) + \sqrt{N}|b^{\gamma_N} - \beta_0| - \sqrt{N}|\hat{b}^{\gamma_N} - b^{\gamma_N}| - \sqrt{N}|\beta_0 - \hat{\beta}| \leq 2\hat{\sigma}\big) \\ &= P\big(\sqrt{N}|b^{\gamma_N} - \beta_0| \leq 2\hat{\sigma} - \hat{\sigma}(\gamma_N) + \sqrt{N}|\hat{b}^{\gamma_N} - b^{\gamma_N}| + \sqrt{N}|\beta_0 - \hat{\beta}|\big) \\ &\leq P\big(\big|\hat{\sigma}(\gamma_N) - 2\hat{\sigma} - \sqrt{N}|\hat{b}^{\gamma_N} - b^{\gamma_N}| - \sqrt{N}|\beta_0 - \hat{\beta}|\big| \geq \sqrt{N}|b^{\gamma_N} - \beta_0|\big) \\ &\leq P\big(|\hat{\sigma}(\gamma_N) - 2\hat{\sigma} - \sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) - \sqrt{N}(\beta_0 - \hat{\beta})| \geq \sqrt{N}|b^{\gamma_N} - \beta_0|\big) \end{aligned}$$

by the reverse triangle inequality. Let us introduce the random variable

$$\Xi_N := \hat{\sigma}(\gamma_N) - 2\hat{\sigma} - \sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) - \sqrt{N}(\beta_0 - \hat{\beta})$$

and the deterministic number $C_N := \sqrt{N}|b^{\gamma_N} - \beta_0|$. By Lemma J.6, we have $\Xi_N = O_P(1)$. Let $\varepsilon > 0$, and choose C_ε and N_ε such that for all $N \geq N_\varepsilon$ the statement $P(|\Xi_N| > C_\varepsilon) < \varepsilon$ holds. By Proposition 4.2, C_N tends to infinity as $N \to \infty$ due to $\gamma_N = o(\sqrt{N})$. Hence, there exists some $\widetilde{N} = \widetilde{N}(C_\varepsilon)$ such that we have $C_N > C_\varepsilon$ for all $N \geq \widetilde{N}$. This implies $P(|\Xi_N| > C_N) \leq P(|\Xi_N| > C_\varepsilon)$ for all $N \geq \widetilde{N}$.

Let $\overline{N} := \max\{N_\varepsilon, \widetilde{N}\}$. For all $N \geq \overline{N}$, we therefore have $P(|\Xi_N| > C_N) < \varepsilon$. We conclude $\lim_{N\to\infty} P(|\Xi_N| > C_N) = 0$. ☐

Lemma J.4. *Let $\gamma_N = o(\sqrt{N})$. We have $\sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) = O_P(1)$.*

Proof of Lemma J.4. We already verified $\hat{D}_1 = D_1 + o_P(1)$ and $\hat{D}_2 = D_2 + o_P(1)$ in the proof of Theorem 4.1. Let us assume that γ_N diverges to $+\infty$ as $N \to \infty$. We then have

$$\begin{aligned}
&\big(\hat{D}_1 + (\gamma_N - 1)\hat{D}_2\big)^{-1} \\
&= \tfrac{1}{\gamma_N - 1}\Big(\tfrac{1}{\gamma_N - 1} D_1 + D_2 + o_P(1) + \tfrac{1}{\gamma_N - 1} o_P(1)\Big)^{-1} \\
&= \tfrac{1}{\gamma_N - 1}\Big(\Big(\tfrac{1}{\gamma_N - 1} D_1 + D_2\Big)^{-1} + o_P(1)\Big) \\
&= \big(D_1 + (\gamma_N - 1)D_2\big)^{-1} + o_P\big(\tfrac{1}{\gamma_N - 1}\big)
\end{aligned}$$

because $\frac{1}{\gamma_N - 1} = O(1)$ holds. Furthermore, we have

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) \\
&= \Big(\big(D_1 + (\gamma_N - 1)D_2\big)^{-1} + o_P\big(\tfrac{1}{\gamma_N - 1}\big)\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \tfrac{1}{\sqrt{n}} \sum_{i \in I_k} \Big(\widetilde{\psi}(S_i; b^{\gamma_N}, \hat{\eta}^{I_k^c}) \\
&\qquad + (\gamma_N - 1)\tfrac{1}{n}\textstyle\sum_{i \in I_k} \psi_1(S_i; \hat{\eta}^{I_k^c})\Big(\tfrac{1}{n}\sum_{i \in I_k} \psi_2(S_i; \hat{\eta}^{I_k^c})\Big)^{-1} \psi(S_i; b^{\gamma_N}, \hat{\eta}^{I_k^c})\Big)
\end{aligned}$$

by (14). Lemma I.16 states that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i \in I_k} \varphi(S_i; b^0, \hat{\eta}^{I_k^c}) - \frac{1}{\sqrt{n}} \sum_{i \in I_k} \varphi(S_i; b^0, \eta^0) \right\| = O_P(\rho_N)$$

holds for $k \in [K]$, $\varphi \in \{\psi, \widetilde{\psi}, \psi_2\}$, and $b^0 \in \{b^\gamma, \beta_0, \mathbf{0}\}$, and where $\rho_N = r_N + N^{\frac{1}{2}}\lambda_N$ is as in Definition I.4 and satisfies $\rho_N \lesssim \delta_N^{\frac{1}{4}}$, and where we interpret $\psi_2(S; b, \eta) = \psi_2(S; \eta)$. This statement remains valid in the present setting because there exists some finite real constant C such that we have $|b^{\gamma_N}| \leq C$ for N large enough. Hence, we have

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) \\
&= \Big(\Big(\tfrac{1}{\gamma_N - 1} D_1 + D_2\Big)^{-1} + o_P(1)\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \Big(\tfrac{1}{\sqrt{n}} \sum_{i \in I_k} \Big(\tfrac{1}{\gamma_N - 1}\widetilde{\psi}(S_i; b^{\gamma_N}, \eta^0) + D_3 \psi(S_i; b^{\gamma_N}, \eta^0) \\
&\qquad + \big(\psi_1(S_i; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)]\big) D_5 - D_3\big(\psi_2(S_i; \eta^0) - \mathbb{E}_P[\psi_2(S; \eta^0)]\big) D_5\Big) \\
&\qquad + o_P(1)\Big)
\end{aligned}$$

by (64). Consider the random variables

$$\begin{aligned}
&\widetilde{X}_i \\
&:= \tfrac{1}{\gamma_N - 1}\widetilde{\psi}(S_i; b^{\gamma_N}, \eta^0) + D_3 \psi(S_i; b^{\gamma_N}, \eta^0) + \big(\psi_1(S_i; \eta^0) - \mathbb{E}_P[\psi_1(S; \eta^0)]\big) D_5 \\
&\quad - D_3\big(\psi_2(S_i; \eta^0) - \mathbb{E}_P[\psi_2(S; \eta^0)]\big) D_5
\end{aligned}$$

for $i \in [N]$, $S_n := \sum_{i \in I_k} \widetilde{X}_i$, and $V_n := \sum_{i \in I_k} \mathbb{E}_P[\widetilde{X}_i^2]$, where $n = \frac{N}{K}$ denotes the size of I_k. The Lyapunov condition is satisfied for $\delta = 2 > 0$ because

$$\frac{1}{\left(\sum_{i \in I_k} \mathbb{E}_P[\widetilde{X}_i^2]\right)^{2+\delta}} \sum_{i \in I_k} \mathbb{E}_P\left[|\widetilde{X}_i|^{2+\delta}\right] = \frac{1}{(\mathbb{E}_P[\widetilde{X}_1^2])^{2+\delta}} \cdot \frac{1}{n^{1+\delta}} \mathbb{E}_P\left[|\widetilde{X}_1|^{2+\delta}\right] \to 0$$

holds as $n \to \infty$. Therefore, the Lindeberg–Feller condition is satisfied, which implies $\frac{S_n}{V_n} \to \mathcal{N}(0,1)$ as $n \to \infty$.

The case where the sequence γ_N is bounded can be analyzed analogously. □

Lemma J.5. *Let $\gamma_N = o(\sqrt{N})$. We then have $\hat{\sigma}^2(\gamma_N) = O_P(1)$.*

Proof of Lemma J.5. We have

$$\hat{\sigma}^2(\gamma_N) = \left(\hat{D}_1 + (\gamma_N - 1)\hat{D}_2\right)^{-1} \hat{D}_4 \left(\hat{D}_1^T + (\gamma_N - 1)\hat{D}_2^T\right)^{-1}.$$

As verified in the proof of Theorem 4.1, we have $\hat{D}_1 = D_1 + o_P(1)$ and $\hat{D}_2 = D_2 + o_P(1)$. We established $\hat{D}_4^k = D_4 + o_P(1)$ in the proof of Theorem J.3 for fixed γ. Consequently, the claim follows if the sequence $\{\gamma_N\}_{N \geq 1}$ is bounded. Next, assume that γ_N diverges to $+\infty$ as $N \to \infty$. We verified

$$\left(\hat{D}_1 + (\gamma_N - 1)\hat{D}_2\right)^{-1} = \left(D_1 + (\gamma_N - 1)D_2\right)^{-1} + o_P\left(\frac{1}{\gamma_N - 1}\right)$$

in the proof of Lemma J.4. It can be shown that $\frac{1}{(\gamma_N - 1)^2}\hat{D}_4$ is bounded in P-probability by adapting the arguments presented in the proof of Theorem J.3 because there exists some finite real constant C such that we have $|b^{\gamma_N}| \leq C$ for N large enough. Therefore,

$$\begin{aligned} &\hat{\sigma}^2(\gamma_N) \\ &= \left(\tfrac{1}{\gamma_N - 1} D_1 + D_2 + o_P(1)\right)^{-1} \tfrac{1}{(\gamma_N - 1)^2} \hat{D}_4 \left(\tfrac{1}{\gamma_N - 1} D_1^T + D_2^T + o_P(1)\right)^{-1} \end{aligned}$$

is bounded in P-probability. □

Lemma J.6. *Let $\gamma = o(\sqrt{N})$. We then have*

$$\Xi_N := \hat{\sigma}(\gamma_N) - 2\hat{\sigma} - \sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) - \sqrt{N}(\beta_0 - \hat{\beta}) = O_P(1).$$

Proof of Lemma J.6. By Theorem 3.1, the term $\sqrt{N}(\beta_0 - \hat{\beta})$ asymptotically follows a Gaussian distribution and is hence bounded in P-probability. By Theorem I.21, the term $\hat{\sigma}^2$ converges in P-probability. Thus, $2\hat{\sigma}$ is bounded in P-probability as well. By Lemma J.4, we have $\sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) = O_P(1)$. By Lemma J.5, we have $\hat{\sigma}^2(\gamma_N) = O_P(1)$. □

Proof of Theorem 4.4. That the statement holds uniformly for $P \in \mathcal{P}_N$ can be derived using analogous arguments as used to prove Theorem 3.1 and 4.1. Theorem J.3 in the appendix shows that $\hat{\sigma}(\gamma)$ consistently estimates $\sigma(\gamma)$ for fixed γ. Analogous arguments show that $\hat{\sigma}(\hat{\gamma}')$ consistently estimates σ from Theorem 3.1. Let $\hat{\mu} := \hat{\gamma}' - 1$. We have

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\hat{\gamma}'} - b^{\hat{\gamma}'}) \\
&= \sqrt{N}\Big(\tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \big(\tfrac{1}{\hat{\mu}}\mathbb{1} + \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\Big)^{-1} \\
&\quad \cdot \tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \big(\tfrac{1}{\hat{\mu}}\mathbb{1} + \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\big(\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} b^{\hat{\gamma}'}\big).
\end{aligned}$$

Due to Theorem 4.3, we have $\frac{1}{\hat{\mu}} = \frac{1}{\sqrt{N}} o_P(1)$. Due to Proposition 4.2, whose statements also hold stochastically for random γ, we have $b^{\hat{\gamma}'} = \beta_0 + \frac{1}{\sqrt{N}} o_P(1)$. Therefore, we have

$$\begin{aligned}
&\sqrt{N}(\hat{b}^{\hat{\gamma}'} - b^{\hat{\gamma}'}) \\
&= \sqrt{N}\Big(\tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\Big)^{-1} \tfrac{1}{K}\textstyle\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \big(\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\beta_0\big) \\
&\quad + o_P(1) \\
&= \sqrt{N}(\hat{\beta} - \beta_0) + o_P(1)
\end{aligned}$$

due to Slutsky's theorem and similar arguments as presented in the proofs of Theorem 3.1 and 4.1. □

Appendix K: Proof of Section 5.1

We argue that A_1 and A_2 are independent of H conditional on W_1 and W_2 in the SEM in Figure 4. First, we consider A_1. All paths from A_1 to H through X or Y are blocked by the empty set because either X or Y is a collider on these paths. The path $A_1 \to A_2 \to W_1 \to H$ is blocked by W_1. Second, we consider A_2. All paths from A_2 to H through X or Y are blocked by the empty set because either X or Y is a collider on these paths. The path $A_2 \to W_1 \to H$ is blocked by W_1.

Acknowledgments

We thank Matthias Löffler, the editor, associate editor, and anonymous reviewers for constructive comments.

References

[1] Acemoglu, D., Johnson, S. and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review* **91** 1369–1401.

[2] Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843.

[3] Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics* **2** 105–110.

[4] Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.

[5] Anderson, T. W. (1983). Some recent developments on the distributions of single-equation estimators. In *Advances in econometrics*, (A. Deaton, D. McFadden and H. Sonnenschein, eds.). *Econometric Society Monographs in Quantitative Economics* 4, 109–122. Cambridge University Press, Cambridge.

[6] Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *Journal of Econometrics* **127** 1–16.

[7] Anderson, T. W., Kunitomo, N. and Sawa, T. (1982). Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator. *Econometrica* **50** 1009–1027.

[8] Anderson, T. W., Kunitomo, N. and Morimune, K. (1986). Comparing single-equation estimators in a simultaneous equation system. *Econometric Theory* **2** 1–32.

[9] Anderson, T. W., Kunitomo, N. and Matsushita, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics* **157** 191–204.

[10] Anderson, T. W. and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics* **20** 46–63.

[11] Anderson, T. W. and Sawa, T. (1979). Evaluation of the Distribution Function of the Two-Stage Least Squares Estimate. *Econometrica* **47** 163–182.

[12] Andrews, I., Stock, J. and Sun, L. (2019). Weak instruments in IV regression: Theory and practice. *Annual Review of Economics* **11** 727–753.

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

[14] Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* **47** 1148–1178.

[15] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972.

[16] Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* **25** 77–83.

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

[18] Berndt, E. R., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3** 653–665.

[19] Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics* **10** 647–671.

[20] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous

Analysis of Lasso and Dantzig Selector. *The Annals of Statistics* **37** 1705–1732.

[21] Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90** 443–450.

[22] Bowden, R. J. and Turkington, D. A. (1985). *Instrumental variables*. *Econometric Society Monographs*. Cambridge University Press, Cambridge.

[23] Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science* **35** 404–426.

[24] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg.

[25] Bühlmann, P. and van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics & Probability Letters* **136** 37–41.

[26] Candes, E. and Tao, T. (2007). The Dantzig Selector: Statistical Estimation When p Is Much Larger than n. *The Annals of Statistics* **35** 2313–2351.

[27] Chen, J., Huang, C.-H. and Tien, J.-J. (2021). Debiased/Double Machine Learning for Instrumental Variable Quantile Regressions. *Econometrics* **9**.

[28] Chen, B., Liang, H. and Zhou, Y. (2016). GMM estimation in partial linear models with endogenous covariates causing an over-identified problem. *Communications in Statistics - Theory and Methods* **45** 3168–3184.

[29] Chernozhukov, V., Hansen, C. and Spindler, M. (2016). hdm: High-dimensional metrics. *R Journal* **8** 185–199.

[30] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Newey, W. and Robins, J. (2017). Repo for the paper "Double/debiased machine learning for treatment and structural parameters". https://github.com/VC2015/DMLonGitHub. Accessed: September 23, 2020.

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

[32] Chiang, H. D., Kato, K., Ma, Y. and Sasaki, Y. (2021). Multiway Cluster Robust Double/Debiased Machine Learning. *Journal of Business & Economic Statistics* **0** 1–11.

[33] Colangelo, K. and Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. Preprint arXiv:2004.03036.

[34] Cragg, J. G. (1967). On the Relative Small-Sample Properties of Several Structural-Equation Estimators. *Econometrica* **35** 89–110.

[35] Crown, W. H., Henk, H. J. and Vanness, D. J. (2011). Some cautions on the use of instrumental variables estimators in outcomes research: How bias in instrumental variables estimators is affected by instrument strength,

instrument contamination, and sample size. *Value in Health* **14** 1078–1084.
[36] Cui, Y. and Tchetgen Tchetgen, E. (2020). Selective machine learning of doubly robust functionals. Preprint arXiv:1911.02029.
[37] DasGupta, A. (2008). *Asymptotic theory of statistics and probability. Springer Texts in Statistics*. Springer, New York.
[38] DiazOrdaz, K., Daniel, R. and Kreif, N. (2019). Data-adaptive doubly robust instrumental variable methods for treatment effect heterogeneity. Preprint arXiv:1802.02821.
[39] Durrett, R. (2010). *Probability: Theory and examples*, 4 ed. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
[40] Emmenegger, C. (2021). dmlalg: Double machine learning algorithms R-package available on CRAN.
[41] Farbmacher, H., Huber, M., Lafférs, L., Langen, H. and Spindler, M. (2020). Causal mediation analysis with double machine learning. Preprint arXiv:2002.12710.
[42] Florens, J.-P., Johannes, J. and Van Bellegem, S. (2012). Instrumental regression in partially linear models. *The Econometrics Journal* **15** 304–324.
[43] Fuller, W. A. (1977). Some Properties of a Modification of the Limited Information Estimator. *Econometrica* **45** 939–53.
[44] Fuller, W. A. (1987). *Measurement error models. Wiley series in probability and mathematical statistics*. John Wiley & Sons, New York.
[45] Hahn, J., Hausman, J. and Kuersteiner, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal* **7** 272–306.
[46] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.
[47] Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* **30** 203–238.
[48] Härdle, W., Liang, H. and Gao, J. (2000). *Partially linear models. Contributions to Statistics*. Springer, Berlin Heidelberg.
[49] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and semiparametric models. Springer series in statistics*. Springer, Berlin.
[50] Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review* **23** 53–60.
[51] Hill, R. C., Griffiths, W. E. and Lim, G. C. (2011). *Principles of econometrics*, 4 ed. Wiley, Hoboken, New Jersey.
[52] Hillier, G. H. and Skeels, C. L. (1993). Some further exact results for structural equation estimators. In *Models, Methods and Applications of Econometrics: essays in Honor of A. R. Bergstroms* (P. C. B. Phillips, ed.) 117–139. Blackwell, Cambridge, Massachusetts.
[53] Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* **79** 347–394.

[54] Jakobsen, M. E. and Peters, J. (2021). Distributional Robustness of K-class Estimators and the PULSE. *The Econometrics Journal*.

[55] Knaus, M. C. (2020). Double machine learning based program evaluation under unconfoundedness. Preprint arXiv:2003.03191.

[56] Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics* **38** 3660–3695.

[57] Kozbur, D. (2020). Analysis of Testing-Based Forward Model Selection. *Econometrica* **88** 2147–2173.

[58] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press, Cambridge.

[59] Lauritzen, S. L. (1996). *Graphical models*. *Oxford statistical science series*. Clarendon Press, Oxford.

[60] Lewis, G. and Syrgkanis, V. (2020). Double/debiased machine learning for dynamic treatment effects. Preprint arXiv:2002.07285.

[61] Liu, M., Zhang, Y. and Zhou, D. (2021). Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*.

[62] Lloyd, W. P. (1975). A Note on the Use of the Two-Stage Least Squares Estimator in Financial Models. *The Journal of Financial and Quantitative Analysis* **10** 143–149.

[63] Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association* **101** 1465–1474.

[64] Maathuis, M., Drton, M., Lauritzen, S. and Wainwright, M., eds. (2019). *Handbook of graphical models*. *Handbooks of Modern Statistical Methods*. Chapman & Hall/CRC, Boca Raton, FL.

[65] Mammen, E. and van de Geer, S. (1997). Penalized Quasi-Likelihood Estimation in Partial Linear Models. *The Annals of Statistics* **25** 1014–1035.

[66] Mariano, R. S. (1972). The Existence of Moments of the Ordinary Least Squares and Two-Stage Least Squares Estimators. *Econometrica* **40** 643–652.

[67] Mariano, R. S. (1982). Analytical Small-Sample Distribution Theory in Econometrics: The Simultaneous-Equations Case. *International Economic Review* **23** 503–533.

[68] Mariano, R. S. (2003). *Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications* In *A Companion to Theoretical Econometrics* 6, 122–141. John Wiley & Sons, Ltd.

[69] Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37** 3779–3821.

[70] Nagar, A. L. (1959). The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* **27** 575–595.

[71] Nagar, A. L. (1960). A Monte Carlo Study of Alternative Simultaneous Equation Estimators. *Econometrica* **28** 573–590.

[72] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, **4** 36, 2111–2245. Elsevier

Science.

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

[74] Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research* **27** 226–284.

[75] Pearl, J. (2004). Robustness of causal claims. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. UAI '04* 446–453. AUAI Press, Arlington, Virginia, USA.

[76] Pearl, J. (2009). *Causality: Models, reasoning, and inference*, 2 ed. Cambridge University Press, Cambridge.

[77] Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics* **6** Article 7.

[78] Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms. Adaptive computation and machine learning*. The MIT Press, Cambridge, MA.

[79] Phillips, P. C. B. (1984). The Exact Distribution of LIML: I. *International Economic Review* **25** 249–261.

[80] Phillips, P. C. B. (1985). The Exact Distribution of LIML: II. *International Economic Review* **26** 21–36.

[81] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954.

[82] Rothenhäusler, D., Meinshausen, N., Bühlmann, P. and Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 215-246.

[83] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression. Cambridge series in statistical and probabilistic mathematics* **12**. Cambridge University Press, Cambridge.

[84] Smucler, E., Rotnitzky, A. and Robins, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. Preprint arXiv:1904.03737.

[85] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **50** 413–436.

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

[87] Stock, J. H., Wright, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* **20** 518–529.

[88] Su, L. and Zhang, Y. (2016). Semiparametric estimation of partially linear dynamic panel data models with fixed effects. In *Essays in Honor of Aman Ullah*, 1 ed. (G. González-Rivera, R. C. Hill and T.-H. Lee, eds.). *Advances in Econometrics* **36** 137–204. Emerald Group Publishing Limited, Howard House, Wagon Lane, Bingley BD16 1WA, UK.

[89] Summers, R. (1965). A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators. *Econometrica* **33** 1–41.

[90] Takeuchi, K. and Morimune, K. (1985). Third-Order Efficiency of the

Extended Maximum Likelihood Estimators in a Simultaneous Equation System. *Econometrica* **53** 177–200.

[91] Theil, H. (1953a). Repeated least-squares applied to complete equation systems. *Central Planning Bureau, The Hague*. Mimeographed memorandum.

[92] Theil, H. (1953b). Estimation and simultaneous correlation in complete equation systems. *Central Planning Bureau, The Hague*. Mimeographed memorandum.

[93] Theil, H. (1961). *Economic forecasts and policy*, 2 ed. *Contributions to economic analysis* **15**. North-Holland Publishing Company, Amsterdam.

[94] van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality. Springer series in statistics*. Springer, New York.

[95] Wager, S. and Walther, G. (2016). Adaptive concentration of regression trees, with application to random forests. Preprint arXiv:1503.06388.

[96] Wagner, H. M. (1958). A Monte Carlo Study of Estimates of Simultaneous Linear Structural Equations. *Econometrica* **26** 117–133.

[97] Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach*, 5 ed. South-Western Cengage Learning, Mason, OH.

[98] Yao, F. (2012). Efficient semiparametric instrumental variable estimation under conditional heteroskedasticity. *Journal of Quantitative Economics* **10** 32–55.

[99] Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high-dimensional additive models. *The Annals of Statistics* **44** 2564–2593.