Diss. ETH No. 29054

Statistical Machine Learning for Complex Data

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zurich)

presented by

CORINNE RAHEL EMMENEGGER

MSc ETH Mathematics, ETH Zurich

born on 09.07.1995

accepted on the recommendation of

Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Rajen Shah, co-examiner

2023

Acknowledgements

I am very happy that I could spend a truly amazing and wonderful time at the Seminar for Statistics at ETH. Thank you all so much! I would like to especially thank my supervisor, Peter Bühlmann, for supporting and guiding me. During my PhD, I had the unique opportunity to work with awesome collaborators who helped me to look at questions from a different angle to come up with creative solutions together: thank you very much David Carl, Timon Elmer, Zijian Guo, Nicolai Meinshausen, Jeffrey Näf, and Meta-Lina Spohn. I would also like to thank my co-supervisor, Rajen Shah, for enlightening discussions.

Abstract

Data are often complex in the sense that they feature dependence between individual observations, unobserved variables, or highly non-linear and interaction effects. For such complex data, we propose algorithms to estimate functionals of interest, like causal treatment effects, linear effects, and conditional distributions. Our first set of methods uses ideas from double machine learning to estimate and make inference for causal treatment effects from observational network data and linear parameters from repeated measurements data and data featuring hidden variables in the presence of high- or infinite-dimensional nuisance components. Our last method is a Random Forest-based algorithm to estimate multivariate conditional distributions.

Zusammenfassung

Daten sind oft komplex in dem Sinne, dass sie Abhängigkeiten zwischen den einzelnen Beobachtungen, unbeobachtete Variablen, oder hochgradig nichtlineare Terme und Interaktionseffekte aufweisen. Für solch komplexe Daten präsentieren wir Algorithmen, um Funktionale von Interesse wie kausale Behandlungseffekte, lineare Effekte und bedingte Verteilungen zu schätzen. Unsere erste Reihe von Methoden verwendet Ideen des doppelten maschinellen Lernens, um kausale Behandlungseffekte aus beobachteten Netzwerkdaten und lineare Parameter aus wiederholten Messdaten und Daten mit nicht observierten Variablen in Gegenwart von hoch- oder unendlich-dimensionalen Störkomponenten zu schätzen. Unsere letzte Methode ist ein Random Forest-basierter Algorithmus zur Schätzung multivariater bedingter Verteilungen.

Contents

1 | Introduction

Data is constantly being collected—be it for example in health care, in education, or in customer behavior analyses. Such data may come from empirical observations instead of carefully designed randomized experiments, from repeatedly observing the same subjects over time, or from simply collecting data without having an analysis goal in mind. Moreover, machine learning algorithms can be key to unlocking the value of such data that feature highly non-linear and interaction terms. However, statistical inference is oftentimes not well understood in these settings. We present methods to estimate and make inference for target functionals of interest from such complex data building on the machine learning tools double machine learning (Chernozhukov et al., 2018) and Distributional Random Forests (Ćevid et al., 2022).

Double machine learning is a tool to estimate and make inference on a low-dimensional parameter θ^0 in the presence of high- or infinite-dimensional nuisance components η^0 that satisfy some moment conditions

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\psi(S_i,\theta^0,\eta^0)]=0,$$

where N denotes the number of experimental units and ψ is a suitable function on the data S_i of the experimental units. Double machine learning uses sample splitting to estimate all nuisance parameters on one part of the data and cross-fitting to build an estimator of θ^0 on the other part of the data by plugging in the nuisance parameter estimates from the first step into the estimating equation. More precisely, the data is partitioned into K many sets $I_1, I_2, \ldots, I_K$ of approximately equal size. For each $k \in \{1, 2, \ldots, K\}$, the nuisance components η^0 are estimated on the complement of I_k using an arbitrary machine learning algorithm and plugged into the estimating equation for θ^0. Then, the data from I_k is used to build an estimator $\hat{\theta}^{I_k}$ of θ^0 using this estimating equation. The final estimator of θ^0 averages over the $\hat{\theta}^{I_k}$, and it converges at the parametric rate, $N^{-1/2}$, and follows a Gaussian distribution asymptotically, provided ψ is Neyman orthogonal and the machine learning errors decay fast enough. Typically, these errors decay at the rate $o_P(N^{-1/4})$ if the problem is sufficiently smooth or sparse. Neyman orthogonality requires that the Gateaux derivative of ψ vanishes at the true θ^0 and η^0, which makes ψ insensitive to inserting biased machine learning estimators of η^0. The algorithm is called "double" machine learning because η^0 consists of at least two objects, which means that machine learning algorithms are applied at least twice. Nonparametric components can also be estimated without sample

splitting (Mammen and van de Geer, 1997), but complex machine learners do not satisfy the entropy conditions these results require (Chernozhukov et al., 2018). Consequently, sample splitting is essential.

This base double machine learning algorithm assumes that unit-level data, S_i, is independent and identically distributed. However, this may not be satisfied in practice. In Chapter 2, we estimate and make inference for a causal treatment effect for observational network data, where experimental units may interact. For example, the vaccination (treatment) of a person not only influences this person's health status (outcome), but can also protect the health status of other people the person is interacting with. Ignoring such interactions may yield biased estimators and invalid inference and contributes to the replication crisis. Nevertheless, practitioners often use data analysis methods that cannot account for such interactions. Our algorithm uses a network, which is an undirected graph on the units, to account for unit-level interactions. We show that the resulting treatment effect estimator is still interpretable in our framework. Our approach uses sample splitting and cross fitting to estimate all nuisance components with machine learning. Compared to existing treatment effect estimators for such a setting, our estimator is easy to implement and asymptotically converges to a Gaussian distribution at the parametric $N^{-1/2}$-rate.

In Chapter 3, we consider repeated measurements data collected on the same units like for instance in a longitudinal trial, which violates the independent and identically distributed assumption. Partially linear mixed-effects models (Zeger and Diggle, 1994; Pinheiro and Bates, 2000) are a powerful tool to cope with such repeated measurements data, but traditional approaches use splines or kernels in combination with parametric estimation to infer the linear coefficient (fixed effects). We propose a machine learning-based approach to obtain a semiparametrically efficient estimator of the fixed effects in the presence of complex interaction structures, nonsmooth terms, and high-dimensional variables.

In Chapter 4, we consider estimating linear effects from data featuring endogeneity; that is, the data feature unobserved variables that correlate the error of the response with the covariates used to explain the response. Although the datapoints are independent across units in this setting, endogeneity leads to unfavorable dependence within individual unit-level data. Two-stage least squares (Theil, 1953a; Basmann, 1957) is a popular tool to cope with endogeneity. Chernozhukov et al. (2018) used their double machine learning framework to do two-stage least squares in partially linear endogenous models to estimate and make inference for the linear model parameter. However, two-stage least squares is known to often produce overly wide confidence intervals in practice. We present a regularization scheme similar to k-class estimators (Theil, 1961)

for linear models that trades off some bias with a reduction in variance, leading to more precise results empirically. Extensive simulation studies complement our theoretical developments on the asymptotic behavior of our estimator.

In Chapter 5, we present our last analysis tool for complex data. We develop uncertainty assessments and confidence intervals with Distributional Random Forests (Ćevid et al., 2022), which is a Random Forest-based (Breiman, 2001) algorithm using Hilbert space embeddings to estimate complex multivariate conditional distributions in a nonparametric way. Furthermore, we discuss two lines of applications of our asymptotic theory: comparing whole distributions of a treatment and a control group to formally identify and test differences between them that may not be captured by the mean alone and estimating functionals of the conditional distribution such as conditional average treatment effects, conditional quantiles, and conditional correlations.

The remaining chapters of this thesis consist of a previously published article or a preprint, up to minor modifications:

Chapter 2: C. Emmenegger, M.-L. Spohn, T. Elmer, and P. Bühlmann. Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning, 2022. Preprint arXiv:2206.14591

Chapter 3: C. Emmenegger and P. Bühlmann. Plugin machine learning for partially linear mixed-effects models with repeated measurements, 2021a. Preprint arXiv:2108.13657

Chapter 4: C. Emmenegger and P. Bühlmann. Regularizing double machine learning in partially linear endogenous models. Electronic Journal of Statistics, 15(2):6461–6543, 2021

Chapter 5: J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Inference for the distributional random forest, 2023. Preprint on arXiv:2302.05761

Chapter 6 at the end of this thesis presents the R-package `dmlalg` (Emmenegger, 2021) that implements the methodology presented in Chapter 3 and 4.

2 | Treatment Effect Estimation from Observational Network Data Using Augmented Inverse Probability Weighting and Machine Learning

Joint work with

Meta-Lina Spohn, Timon Elmer, and Peter Bühlmann

This chapter is based on the manuscript

C. Emmenegger, M.-L. Spohn, T. Elmer, and P. Bühlmann. Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning, 2022. Preprint arXiv:2206.14591

Abstract

Causal inference methods for treatment effect estimation usually assume independent experimental units. However, this assumption is often questionable because experimental units may interact. We develop augmented inverse probability weighting (AIPW) for estimation and inference of causal treatment effects on dependent observational data. Our framework covers very general cases of spillover effects induced by units interacting in networks. We use plugin machine learning to estimate infinite-dimensional nuisance components leading to a consistent treatment effect estimator that converges at the parametric rate and asymptotically follows a Gaussian distribution. We apply our AIPW method to the Swiss StudentLife Study data to investigate the effect of hours spent studying on exam performance accounting for the students' social network.

2.1 | Introduction

Classical causal inference from observational data usually assumes that the experimental units are independent. This assumption is also part of the popular stable unit treatment value assumption (SUTVA) (Rubin, 1980). However, independence is often not conceivable in practice due to interactions among

units, and so-called spillover effects may occur. For example, participants in a clinical trial on an infectious disease may interact, and the vaccination (treatment) of a person not only influences this person's health status (outcome), but can also protect the health status of other people the person is interacting with (Perez-Heydrich et al., 2014; Sävje et al., 2021). Alternatively, in a housing mobility experiment, a random selection of residents receive information to assist their relocation, but such additional information might also influence the behavior of people with whom the recipients communicate (Sobel, 2006). Other studies where dependencies come from spillover effects include reducing students' depressive symptoms (Vanderweele et al., 2013), teaching methods in education (Hong and Raudenbush, 2008), school-based deworming programs (Miguel and Kremer, 2003; Aiken et al., 2015), and fMRI imaging for motor inhibition (Luo et al., 2012).

Causal inference in the presence of spillover effects is an enormous challenge. We no longer have N independent observed realizations to learn relevant properties of the underlying data generating mechanism. Instead, we only observe a single draw of N dependent units from the data generating mechanism. In the presence of spillover effects, standard algorithms fail to separate correlation from causation, and spurious associations due to network dependence contribute to the replication crisis (Lee and Ogburn, 2021). Falsely relying on assumptions like SUTVA may yield biased causal effect estimators and invalid causal inference; see for instance Sobel (2006), Perez-Heydrich et al. (2014), Ogburn and VanderWeele (2017), Ogburn et al. (2022), Eckles and Bakshy (2021), and Lee and Ogburn (2021). New tailored methods are required to guarantee valid causal inference from observational data with spillover effects. However, there is no established general methodological framework for the latter task, and at least two reasons are associated with this. First, there are numerous natural notions of causal effects, most of them being contrasts of low-dimensional summaries of the counterfactual treatment distributions. Second, additional assumptions are required to describe and control for different spillover effects.

In this paper, the causal effect of interest and target of inference is the so-called expected average treatment effect (EATE) (Sävje et al., 2021). The EATE measures how, on average, the outcome of a unit is causally affected by its own treatment in the presence of spillover effects from other units. For a dichotomous treatment $W_i \in \{0, 1\}$ and an outcome Y_i for units $i = 1, 2, \ldots, N$, the EATE is given by

$$\theta_N^0 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)} \right],$$

where we use the do-notation of Pearl (1995), and the expectation is with respect to all random components, whose distributions are given by the data generating mechanism in (4.3) below. This notation makes explicit that, given a unit i, we

consider the causal effect of the treatment W_i to be the (unobservable) expected difference in outcomes Y_i when the treatment is assigned to unit i versus when the treatment is retained from unit i. The unit-specific spillover effects are not explicitly visible in the EATE because we take the expectation over them. In the infectious disease example given above, the EATE measures the expected difference in health status when an individual receives the vaccination versus when it does not, marginalizing over individual-specific covariates and spillover effects of the other study participants. This corresponds to the medical effect of the vaccine in a person's body, which reflects the direct effect of the treatment (Sävje et al., 2021). Other authors considering the EATE or the similar average direct effect (ADE) (Hudgens and Halloran, 2008) include VanderWeele and Tchetgen Tchetgen (2011), Sofrygin and van der Laan (2017), Sävje et al. (2021), Hu et al. (2022), and Li and Wager (2022). If spillover effects are absent, the EATE matches the expected value of the usual average treatment effect (Splawa-Neyman et al., 1990; Rubin, 1974).

We consider the following types of spillover effects: causal effects of other units' treatments on a given unit's outcome, called interference (Sobel, 2006; Hudgens and Halloran, 2008), and causal effects of other units' covariates and confounders on a given unit's treatment or its outcome[1]. To characterize these spillover effects, we assume a known undirected network among the N units, in which the N nodes represent the units and the edges represent some kind of interaction or relationship of the respective units such as friendship, geographical closeness, or shared department in a company. We then use features that are arbitrary functions of this network and the treatment and covariate vectors of the whole population (Manski, 1993; Chin, 2019). The features are assumed to capture all pathways through which spillover effects take place and are specified by the user. For example, Cai et al. (2015) and Leung (2020) model the purchase of a weather insurance (outcome) of farmers in rural China as a function of attending a training session (treatment) and the proportion of friends (feature on direct neighbors in network) who attend the session.

We consider a structural equation model (SEM) to specify the data generating mechanism. Such an SEM approach is also used by van der Laan (2014), Ogburn et al. (2022), Sofrygin and van der Laan (2017), and Spohn et al. (2023). For simplicity, we consider continuous outcomes in this exposition. Please see Section 2.2.1 for more details. The unit-level observations for $i = 1, 2, \ldots, N$

[1]Another notion of spillover effects is frequently used in the social sciences; please see Section 2.B in the appendix for a discussion.

come from sequentially evaluating the structural equations

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i} \\ W_i &\leftarrow \text{Bernoulli}\big(h^0(C_i, Z_i)\big) \\ Y_i &\leftarrow W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i} \end{aligned}$$

for independent and identically distributed ε_{C_i} and independent and identically distributed and centered ε_{Y_i} that are also independent of the ε_{C_i}'s. For each unit i, we observe the binary treatment $W_i \in \{0, 1\}$, a vector of observed confounders C_i, and an outcome Y_i. The features Z_i (functions of other units' covariates) and X_i (functions of other units' covariates and treatments) denote the spillover effects unit i receives from other units; see Section 2.2.1 for their construction. The functions g_1^0 and g_0^0, which govern the outcome model, and the propensity function h^0 may be highly complex and nonsmooth and include interactions and high-dimensional variables.

We follow an augmented inverse probability weighting (AIPW) (Robins et al., 1995) approach to estimate the EATE θ_N^0 in the context of this model. Inverse probability weighting (IPW) (Rosenbaum, 1987) is used to estimate treatment effects under SUTVA. Under SUTVA, the AIPW approach has reduced variability and improved efficiency compared to IPW. In our setting, we do not restrict to SUTVA due to the non-iid nature of the data, and we estimate g_1^0, g_0^0 and h^0 with arbitrary machine learning algorithms and plug them into our AIPW estimand identifying θ_N^0. If the treatment is randomized, we do not have to estimate the propensity function h^0 and set it to the randomization probability instead. The estimators of g_0^0, g_1^0, and h^0 may be biased, especially if regularization methods are used, like for instance with Lasso (Tibshirani, 1996). However, this bias is absorbed by the estimating equation for θ_N^0 because it is Neyman orthogonal (Chernozhukov et al., 2018). We use the ideas of sample splitting with cross-fitting introduced in the double machine learning framework originally proposed for independent and identically distributed data by Chernozhukov et al. (2018). Our estimator of the EATE converges at the parametric rate, $N^{-1/2}$, and asymptotically follows a Gaussian distribution. This allows us to construct confidence intervals and p-values.

2.1.1 | Our Contribution and Comparison to Literature

Our contribution is five-fold. First, we extend the allowed complexity of the model. We do not require observations from multiple independent groups of units, a randomized treatment, or any sort of parameterization; we refer to Ogburn et al. (2022) for an overview of such "standard" approaches. Second, we present a nonparametric, machine-learning-based approach to estimate the EATE from observational network data that enables performing inference, including confidence intervals and p-values. Third, the limiting asymptotic

Gaussian distribution and optimal $1/\sqrt{N}$ convergence rate of the EATE estimator are achieved even if the spillover effects are not limited to neighboring units in the network and the number of ties of a unit may diverge asymptotically. Fourth, our algorithm is easy to understand and implement. Fifth, we analyze the Swiss StudentLife Study data (Stadtfeld et al., 2019; Vörös et al., 2021) and estimate the effect of studying time on the grade point average of freshmen students after their first-year examinations at one of the world's leading universities.

In contrast, the current literature on non- and semiparametric estimation of causal effects from observational network data consider the following. Liu et al. (2019) propose a parametric and doubly robust estimator of a variety of causal effects under the assumption of observing multiple independent groups of units. Tchetgen Tchetgen et al. (2021) develop a network version of the g-formula (Robins, 1986) and perform outcome regression, assuming that the data can be represented as a chain graph, which is a graphical model that is generally incompatible with our SEM approach (Lauritzen and Richardson, 2002). Tchetgen Tchetgen et al. (2021), van der Laan (2014), Ogburn et al. (2022), and Sofrygin and van der Laan (2017) also develop asymptotic Gaussian theory for causal effect estimation on arbitrary networks. The latter three works consider semiparametric estimation and use targeted maximum likelihood (TMLE) methodology (van der Laan and Rubin, 2006; van der Laan and Rose, 2011, 2018). van der Laan (2014) and Ogburn et al. (2022) primarily consider global effects such as the global average treatment effect (GATE) that contrasts the hypothetical intervention of treating all units in the population versus treating no unit of the population. Sofrygin and van der Laan (2017) mention a possible extension to estimate direct effects as we do, but all their results are for global effects such as the GATE. Furthermore, the TMLE framework requires density estimation, which can be awkward in practice, and the theory assumes some kind of a bounded entropy integral, which typically rules out many modern machine learning methods. Our algorithm is easy to understand and implement, and the user may choose any machine learning algorithm they like. Finally, to achieve the $1/\sqrt{N}$ convergence rate, Sofrygin and van der Laan (2017) uniformly limit the number of neighbors of each unit in the network, and spillover effects are limited to direct neighbors. With our algorithm, the number of interactions of a unit may increase with the sample size, and interactions may be beyond direct network neighbors. Spohn et al. (2023) consider a similar setting as we do, but they focus on graphical identification of causal effects, and their outcome equation is entirely linear.

In randomized experiments, Sävje et al. (2021) establish that the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the Hájek estimator (Hájek, 1971) of the EATE, which were initially designed for no-interference

settings, remain consistent under interference. However, their convergence rate depends on the degree of interference. In randomized experiments, Li and Wager (2022) recover the parametric convergence rate of these estimators and provide a central limit theorem. They consider arbitrary networks, but interference is only due to the average of treated neighbors. Their Horvitz-Thompson and the Hájek estimator do not account for observed confounding.

2.1.2 | Additional Literature

Spillover effects can be of various forms. We consider interference and spillover effects from covariates. Another widespread spillover effect is contagion where the outcome of a unit potentially affects the outcome of other units (Ugander et al., 2013; Eckles et al., 2017). In certain cases, contagion can be expressed as a sum of interference effects (Chin, 2019).

A review of causal inference under interference can be found in Halloran and Hudgens (2016). Sävje et al. (2021) give a comprehensive overview of the development of the network interference literature. Many approaches consider disjoint independent groups and arbitrary interference within groups, an assumption called partial interference (Sobel, 2006). Recent publications on observational data from networks use the potential outcomes framework; see for example Forastiere et al. (2021), Toulis et al. (2021), and Wang (2021).

Also other works consider an SEM framework as we do to model dependent data in the context of networks; see for instance Shalizi and Thomas (2011); Ogburn and VanderWeele (2014, 2017); Taylor and Eckles (2018); Egami and Tchetgen Tchetgen (2021). The corresponding causal directed acyclic graph (DAG) of the data generating mechanism contains the variables of all units, and the observed data are one or multiple observations from such graphs. However, such graphs are highly complex, which may limit the practicality of these approaches. Moreover, if the data is modeled as one realization from the network, statistical inference based on asymptotics may not be possible; see for example Tchetgen Tchetgen et al. (2021). Spohn et al. (2023) and Zhang et al. (2022) study graphical identification of effects with DAGs that are less complex.

Ogburn et al. (2022) and Hoshino (2021) also consider unobserved confounding. If the network is not accurately specified, not recorded edges in the network introduce unobserved confounding. We assume a known network.

Outline of the Paper. Section 2.2 presents the model assumptions, characterizes the treatment effect of interest, outlines the procedures for the point estimation of the EATE and estimation of its variance, and establishes asymptotic results. Section 2.3 demonstrates our methodological and theoretical developments in a simulation study and on empirical data. We investigate the effect of hours spent studying on exam performance in the Swiss StudentLife Study taking into account the effect of social ties.

2.2 | Model Formulation and our Network AIPW Estimator

2.2.1 | Model Formulation

We consider N units for which we observe a binary treatment $W_i \in \{0, 1\}$, a univariate outcome Y_i, and a possibly multivariate vector of observed confounders C_i that may causally affect W_i and Y_i. The outcome Y_i may be dichotomous or continuous, and the confounders C_i may consist of discrete and continuous components. If the outcomes are continuous, the unit-level observations $i = 1, 2, \ldots, N$ are realizations from sequentially evaluating the structural equations

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i} \\ W_i &\leftarrow \text{Bernoulli}\big(h^0(C_i, Z_i)\big) \\ Y_i &\leftarrow W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i} \end{aligned} \tag{2.1}$$

for independent and identically distributed ε_{C_i} and independent and identically distributed and centered ε_{Y_i} that are also independent of the ε_{C_i}'s, and where the spillover features Z_i (functions of other units' confounders) and X_i (functions of other units' confounders and treatments) are defined below. For dichotomous responses, we consider for each unit i the structural equations

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i} \\ W_i &\leftarrow \text{Bernoulli}\big(h^0(C_i, Z_i)\big) \\ Y_i &\leftarrow \text{Bernoulli}\big(W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i)\big) \end{aligned} \tag{2.2}$$

for independent and identically distributed ε_{C_i}. Both sets of structural equations, (2.1) and (2.2), can be represented with additive errors as

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i} \\ W_i &\leftarrow h^0(C_i, Z_i) + \varepsilon_{W_i} \\ Y_i &\leftarrow W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i} \end{aligned} \tag{2.3}$$

for some error terms ε_{W_i} and ε_{Y_i} that are independent across units and satisfy $\mathbb{E}[\varepsilon_{W_i} | C_i, Z_i] = 0$ and $\mathbb{E}[\varepsilon_{Y_i} | W_i, C_i, X_i] = 0$ within units. We will make use of this latter representation. Particularly, the confounders C_i and spillover effects X_i may affect the outcome Y_i of unit i differently depending on its own treatment, captured by potentially different g_1^0 and g_0^0. The features X_i and Z_i capture spillover effects across units along the paths of the underlying known network $G = (V, E)$ and render the unit-level data dependent. The vertex set V of G consists of the N units, and the edge set E consists of pairwise,

undirected edges. The feature X_i is given by the vector

$$X_i = \Big(f_x^1\big(\{(W_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big), \ldots, f_x^r\big(\{(W_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big) \Big)$$

of fixed dimension r, where $[N]$ denotes the set $\{1, 2, \ldots, N\}$. Each function $f_x^l \colon \mathbb{R}^{(N-1)\times(N-1)\times N\times N^2} \to \mathbb{R}$ for $l \in [r]$ is a function of the treatment and confounder vector of units other than i and the network G. Each such function is specified by the user and describes a 1-dimensional spillover effect that unit i receives from other units' confounders and treatments. The feature Z_i is defined analogously as the vector

$$Z_i = \Big(f_z^1\big(\{C_j\}_{j\in[N]\setminus\{i\}}, G\big), \ldots, f_z^t\big(\{C_j\}_{j\in[N]\setminus\{i\}}, G\big) \Big)$$

of fixed dimension t, where each $f_z^l \colon \mathbb{R}^{(N-1)\times N\times N^2} \to \mathbb{R}$ for $l \in [t]$ is a function of the confounder vector of units other than i and the network G. Each such function is specified by the user and describes a 1-dimensional spillover effect that the treatment assignment W_i of unit i receives from other units' confounders. Such an approach was initially proposed by Chin (2019). The functions f_x^l, $l \in [r]$ and f_z^l, $l \in [t]$ are shared by all units and are of fixed dimension. Thus, they help to reduce the dimension and complexity arising from a potentially growing number of influencing units because they map spillover effects to vectors of fixed dimensions. The dependencies captured by the X- and Z-features are reciprocal for two connected units in the undirected network G: if there is an edge between two units i and j, then unit i receives spillover effects from W_j and/or C_j, and unit j receives spillover effects from W_i and/or C_i. Example 2.2.1 illustrates the construction of an X-feature that accounts for treatment spillover from neighbors and neighbors of neighbors. However, also more complex effects additionally accounting for covariate spillovers are possible. As the number of units N increases, the number of connections of each unit may increase (or decrease) as well. Consequently, a unit may receive spillover effects from more (or less) other units as N increases.

We denote the direct causes of W_i by $\mathrm{pa}(W_i)$, the parents of W_i. Analogously, we denote the parents of Y_i by $\mathrm{pa}(Y_i)$; please see for instance Lauritzen (1996). We assume that $\mathrm{pa}(W_i)$ consists of C_i and the variables used to compute the spillover feature Z_i and that $\mathrm{pa}(Y_i)$ consists of W_i, C_i, and the variables used to compute the spillover feature X_i.

Example 2.2.1. *Consider the network in Figure 2.1 where gray nodes receive the treatment and white ones do not. We choose $r = 2$ many X-features and discard any influence of C_j in X_i, that is, we consider the case $f_x^l\big(\{(W_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big) = f_x^l\big(\{(W_j)\}_{j\in[N]\setminus\{i\}}, G\big)$ for $l = 1, 2$. Given*

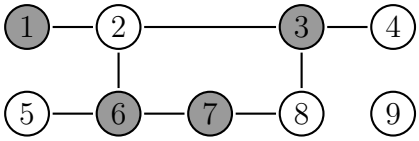


Figure 2.1: A network on nine units where the node label represents the number of a unit. Gray nodes receive the treatment, corresponding to $W_i = 1$, and white ones do not, corresponding to $W_i = 0$.

a unit i, we choose the first feature in X_i as the average number of treated neighbors of unit i and the second feature as the average number of the treated neighbors of neighbors of i. Let us consider unit $i = 6$ in Figure 2.1. Its neighbors are the units 2, 5, and 7, and its neighbors of neighbors are the units 1 and 3 (neighbors of unit 2) and unit 8 (neighbor of unit 7), where we exclude $i = 6$ from its second degree neighborhood by definition. Therefore, we have $X_6 = (1/3, 2/3)$ because one out of three neighbors is treated and two out of three neighbors of neighbors are treated. The whole 9×2 dimensional X-feature matrix is obtained by applying the same computations to all other units i.

2.2.2 | Treatment Effect and Identification

Let us recall the treatment effect of interest, the expected average treatment effect (EATE) (Sävje et al., 2021),

$$\theta_N^0 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)} \right],$$

and let us denote the unit-level direct effect for $i \in [N]$ by

$$\theta_i^0 = \mathbb{E}\left[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)} \right] = \mathbb{E}\left[g_1^0(C_i, X_i) - g_0^0(C_i, X_i) \right],$$

where the second equality comes from inserting the model (4.3). The unit-level direct effect measures how the outcome Y_i of unit i is causally affected by its own treatment assignment W_i in the presence of spillover effects from other units. Consequently, the EATE is given by

$$\theta_N^0 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[g_1^0(C_i, X_i) - g_0^0(C_i, X_i) \right] = \frac{1}{N} \sum_{i=1}^{N} \theta_i^0, \tag{2.4}$$

which is the average of the N unit-specific direct effects. Estimating g_1^0 and g_0^0 by regression machine learning algorithms and plugging them into (2.4) would not result in a parametric convergence rate and an asymptotic Gaussian distribution

of our estimator. To obtain asymptotic normality with convergence at the parametric rate, we add a centered correction term to $g_1^0(C_i, X_i) - g_0^0(C_i, X_i)$, which yields the estimating equation

$$\begin{aligned}\varphi(S_i, \eta) \;=\; & g_1(C_i, X_i) - g_0(C_i, X_i) + \tfrac{W_i}{h(C_i, Z_i)}\big(Y_i - g_1(C_i, X_i)\big) \\ & - \tfrac{1-W_i}{1-h(C_i, Z_i)}\big(Y_i - g_0(C_i, X_i)\big)\end{aligned} \tag{2.5}$$

for θ_N^0, where the unit-level data

$$S_i = (W_i, C_i, X_i, Z_i, Y_i) \tag{2.6}$$

concatenates the variables observed for unit i, and $\eta = (g_1, g_0, h)$ concatenates general nuisance functions g_1, g_0, and h. The true nuisance functions $\eta^0 = (g_1^0, g_0^0, h^0)$ are not of statistical interest to us, but have to be estimated to build an estimator for θ_N^0. The following lemma identifies the EATE, where φ is evaluated at S_i and η^0.

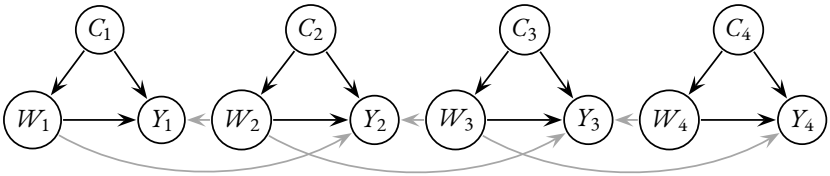
Lemma 2.2.2. *Let $i \in [N]$. It holds that $\mathbb{E}[\varphi(S_i, \eta^0)] = \theta_i^0$, and we can consequently identify the EATE* (2.4) *by*

$$\theta_N^0 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\varphi(S_i, \eta^0)\right]. \tag{2.7}$$

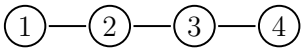
Based on this lemma, we will present our estimator of θ_N^0 in Section 2.2.4. We will estimate g_1^0, g_0^0, and the propensity function h^0 using any regression machine learning algorithms that are allowed to be biased. However, the two correction terms $W_i/h(C_i, Z_i)(Y_i - g_1(C_i, X_i))$ and $(1-W_i)/(1-h(C_i, Z_i))(Y_i - g_0(C_i, X_i))$ make φ Neyman orthogonal and thus insensitive to inserting potentially biased machine learning estimators. Neyman orthogonality is an essential tool to obtain the $1/\sqrt{N}$ convergence rate of the treatment effect estimator; please see Section 2.2.4 for further details.

Scharfstein and Robins (1999) and Bang and Robins (2005) consider a similar function φ for causal effect estimation and inference under the SUTVA assumption, and their function is based on the influence function for the mean for missing data from Robins and Rotnitzky (1995). Moreover, it is also used to compute the AIPW estimator under SUTVA, and our estimating equation φ defined in (2.5) coincides with the one of the AIPW approach under SUTVA if we omit the X- and Z-spillover features. In this case, we can reformulate φ as

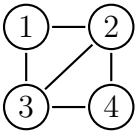
$$\begin{aligned}&\varphi(S_i, \eta^0) \\ =\; & \tfrac{W_i Y_i}{e(C_i)} - \tfrac{(1-W_i)Y_i}{(1-e(C_i))} \\ & - \tfrac{W_i - e(C_i)}{e(C_i)(1-e(C_i))}\Big((1-e(C_i))\,\mathbb{E}[Y_i|W_i=1, C_i] + e(C_i)\,\mathbb{E}[Y_i|W_i=0, C_i]\Big),\end{aligned}$$



(a) Causal DAG



(b) Network G



(c) Dependency graph G_D

Figure 2.2: A network G on four units in 2.2b, where the spillover effects come from the treatments of the direct neighbors, which results in a distance-two dependence, which is displayed in the corresponding dependency graph G_D in 2.2c. The underlying causal DAG is displayed in 2.2a, where arrows due to X-spillover effects are gray.

where $e(C_i) = \mathbb{E}[W_i|C_i] = h^0(C_i)$ denotes the propensity score, $\mathbb{E}[Y_i|W_i = 1, C_i] = g_1^0(C_i, X_i)$, and $\mathbb{E}[Y_i|W_i = 0, C_i] = g_0^0(C_i, X_i)$. This equivalence remains true if the true nuisance functions are replaced by their estimators.

2.2.3 | Dependency Graph

So far, we characterized spillover effects and resulting dependencies among the units by a network. If an edge connects two units, the units may be dependent. However, the absence of an edge in the network does not necessarily imply independence of the respective units. Subsequently, we present a second graph where the presence of an edge represents dependence and its absence independence. Our theoretical results will be established based on this so-called dependency graph (Sävje et al., 2021). Example 2.2.4 illustrates the concept.

Definition 2.2.3 (Dependency graph on S_i, $i \in [N]$)**.** *(Sävje et al., 2021). The dependency graph $G_D = (V, E_D)$ on the unit-level data S_i, $i \in [N]$ defined in* (2.6) *is an undirected graph on the node set V of the network $G = (V, E)$ with potentially larger edge set E_D than E. An undirected edge $\{i, j\}$ between two nodes i and j from V belongs to E_D if at least one of the following two conditions holds:* 1) *there exists an $m \in [N] \setminus \{i, j\}$ such that W_m and/or C_m are present in both X_i and X_j or are present in both Z_i and Z_j;* 2) *W_i is present in X_j, or C_i is present in X_j or in Z_j. That is, units i and j receive spillover effects from at least one common*

confounding unit, or they receive spillover effects from each other.

Example 2.2.4. *Consider the chain-shaped network G in Figure 2.2b. We consider a 1-dimensional X-spillover effect as the fraction of treated direct neighbors in the network G and no Z-spillover. The resulting dependency graph G_D is displayed in Figure 2.2c. In G_D, unit 2 shares an edge with units 1 and 3 because these units are neighbors of 2 in the network. Unit 2 also shares an edge with 4 in G_D because it shares its neighbor 3 with unit 4. Figure 2.2a displays the causal DAG on all units corresponding to this model, including confounders C. Due to the definition of the X-spillover effect, we have $X_1 = W_2$ and $X_4 = W_3$. Consequently, using graphical criteria (Lauritzen, 1996; Pearl, 1998, 2009, 2010; Perković et al., 2018), we infer that the unit-level data $S_1 = (W_1, C_1, X_1, Y_1)$ is independent of $S_4 = (W_4, C_4, X_4, Y_4)$.*

2.2.4 | Estimation Procedure and Asymptotics

Subsequently, we describe our estimation procedure and its asymptotic properties. We use sample splitting and cross-fitting to estimate the EATE θ_N^0 identified by Equation (2.7) as follows. We randomly partition $[N]$ into $K \geq 2$ sets of approximately equal size that we call $I_1, \ldots, I_K$. We split the unit-level data according to this partition into the sets $\mathcal{S}_{I_k} = \{S_i\}_{i \in I_k}$, $k \in [K]$. For each $k \in [K]$, we perform the following steps. First, we estimate the nuisance functions g_1^0, g_0^0, and h^0 on the complement set of $\mathcal{S}_{I_k}$, which we define as

$$\mathcal{S}_{I_k^c} = \{S_j\}_{j \in [N]} \setminus \big(\mathcal{S}_{I_k} \cup \{S_m \mid \exists i \in I_k \colon (i,m) \in E_D\}\big), \tag{2.8}$$

where E_D denotes the edge set of the dependency graph G_D. Particularly, $\mathcal{S}_{I_k^c}$ consists of unit-level data S_j from units j that do not share an edge with any unit $i \in I_k$ in the dependency graph. Consequently, the set $\mathcal{S}_{I_k^c}$ contains all S_j's that are independent of the data in $\mathcal{S}_{I_k}$. To estimate g_1^0, we select the S_i's from $\mathcal{S}_{I_k^c}$ whose W_i equals 1 and regress the corresponding outcomes Y_i on the confounders C_i and the features X_i, which yields the estimator $\hat{g}_1^{I_k^c}$. Similarly, to estimate g_0^0, we select the S_i's from $\mathcal{S}_{I_k^c}$ whose W_i equals 0 and perform an analogous regression, which yields the estimator $\hat{g}_0^{I_k^c}$. To estimate h^0, we use the whole set $\mathcal{S}_{I_k^c}$ and regress W_i on the confounders C_i and the features Z_i, which yields the estimator $\hat{h}^{I_k^c}$. These regressions may be carried out with any machine learning algorithm. We concatenate these nuisance function estimators into the nuisance parameter estimator $\hat{\eta}^{I_k^c} = (\hat{g}_1^{I_k^c}, \hat{g}_0^{I_k^c}, \hat{h}^{I_k^c})$ and plug it into φ that is defined in (2.5). We then evaluate the so-obtained function $\varphi(\cdot, \hat{\eta}^{I_k^c})$ on the data $\mathcal{S}_{I_k}$, which yields the terms $\varphi(S_i, \hat{\eta}^{I_k^c})$ for $i \in I_k$. That is, we evaluate $\varphi(\cdot, \hat{\eta}^{I_k^c})$ on unit-level data S_i that is independent of the data that was used to

estimate the nuisance parameter $\hat{\eta}^{I_k^c}$. Finally, we estimate the EATE by the cross-fitting estimator

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{|I_k|} \sum_{i \in I_k} \varphi(S_i, \hat{\eta}^{I_k^c}) \right) \tag{2.9}$$

that averages over all K folds. The estimator $\hat{\theta}$ converges at the parametric rate, $N^{-1/2}$, and follows a Gaussian distribution asymptotically with limiting variance σ_∞^2 as stated in Theorem 2.2.5 below.

The partition $I_1, \ldots, I_K$ is random. To alleviate the effect of this randomness, the whole procedure is repeated a number of B times, and the median of the individual point estimators over the B repetitions is our final estimator of θ_N^0. The asymptotic results for this median estimator remain the same as for $\hat{\theta}$; see Chernozhukov et al. (2018). For each repetition $b \in [B]$, we compute a point estimator $\hat{\theta}_b$, a variance estimator $\hat{\sigma}_{\infty,b}^2$ (for details please see the next Section 2.2.5), and a p-value p_b for the two-sided test $H_0 \colon \theta_N^0 = 0$ versus $H_A \colon \theta_N^0 \neq 0$. The B many p-values $p_1, \ldots, p_B$ from the individual repetitions are aggregated according to

$$p_{\text{aggr}}^0 = 2 \,\text{median}_{b \in [B]}(p_b).$$

This aggregation scheme yields a valid overall p-value for the same two-sided test (Meinshausen et al., 2009). The corresponding confidence interval is constructed as

$$\text{CI}(\hat{\theta}) = \{\theta \in \mathbb{R} \mid p_{\text{aggr}}^\theta \text{ of } H_0 \colon \theta_N^0 = \theta \text{ vs. } H_A \colon \theta_N^0 \neq \theta \text{ satisfies } p_{\text{aggr}}^\theta > \alpha\}, \tag{2.10}$$

where typically $\alpha = 0.05$. This set contains all values θ for which the null hypothesis $H_0 \colon \theta_N^0 = \theta$ cannot be rejected at level α against the two-sided alternative $H_A \colon \theta_N^0 \neq \theta$.

Next, we describe how $\text{CI}(\hat{\theta})$ can easily be computed. Due to the asymptotic result of Theorem 2.2.5, the aggregated p-value p_{aggr}^θ for $\theta \in \mathbb{R}$ can be represented as

$$p_{\text{aggr}}^\theta = 4 \,\text{median}_{b \in [B]} \left(1 - \Phi(\sqrt{N} \hat{\sigma}_{\infty,b}^{-1} |\hat{\theta}_b - \theta|)\right),$$

where Φ denotes the cumulative distribution function of a standard Gaussian random variable. Consequently, we have

$$p_{\text{aggr}}^\theta > \alpha \quad \Leftrightarrow \quad \Phi^{-1}(1 - \alpha/4) > \text{median}_{b \in [B]}(\sqrt{N} \hat{\sigma}_{\infty,b}^{-1} |\hat{\theta}_b - \theta|),$$

which can be solved for feasible values of θ using root search. A full description of our method is presented in Algorithm 3 in the next Section 2.2.5 after the

description of the variance estimator $\hat{\sigma}_{\infty,b}$.

We now present our main theorem that we mentioned in the construction of confidence intervals above:

Theorem 2.2.5 (Asymptotic distribution of $\hat{\theta}$)**.** *Assume Assumption 3.B.2, 2.A.2, 2.A.3 and 3.B.4 stated in the appendix in Section 3.B. Then, the estimator $\hat{\theta}$ of the EATE θ_N^0 given in* (2.9) *converges at the parametric rate, $N^{-1/2}$, and asymptotically follows a Gaussian distribution, namely*

$$\sqrt{N}(\hat{\theta} - \theta_N^0) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2), \tag{2.11}$$

where σ_∞ is characterized in Assumption 2.A.3. The convergence in (2.11) *is in fact uniformly over the law P of the observations.*

Please see Section 2.E in the appendix for a proof of Theorem 2.2.5. The asymptotic variance σ_∞^2 in Theorem 2.2.5 can be consistently estimated; see Theorem 2.2.6 in the next Section 2.2.5.

Subsequently, we describe the implications of the assumptions made in Theorem 2.2.5. Assumption 2.A.3 ensures that σ_∞^2 exists. Assumption 3.B.2 and 3.B.4 specify regularity conditions and required convergence rates of the machine learning estimators. The machine learning errors need to satisfy the product relationship

$$\begin{aligned}\|h^0(C_i, Z_i) - \hat{h}^{I_k^c}(C_i, Z_i)\|_{P,2} \cdot \Big(&\|g_1^0(C_i, X_i) - \hat{g}_1^{I_k^c}(C_i, X_i)\|_{P,2} \\ &+ \|g_0^0(C_i, X_i) - \hat{g}_0^{I_k^c}(C_i, X_i)\|_{P,2} + \|h^0(C_i, Z_i) - \hat{h}^{I_k^c}(C_i, Z_i)\|_{P,2} \Big) \ll N^{-\frac{1}{2}}.\end{aligned}$$

This bound requires that only products of the machine learner's errors but not the individual ones need to vanish at a rate smaller than $N^{-1/2}$. In particular, the individual error terms may vanish at a rate smaller than $N^{-1/4}$. This is achieved by many machine learning methods; see for instance Chernozhukov et al. (2018): ℓ_1-penalized and related methods in a variety of sparse models (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Belloni et al., 2011; Belloni and Chernozhukov, 2011; Belloni et al., 2012; Belloni and Chernozhukov, 2013), forward selection in sparse models (Kozbur, 2020), L_2-boosting in sparse linear models (Luo and Spindler, 2016), a class of regression trees and random forests (Wager and Walther, 2016), and neural networks (Chen and White, 1999).

We note that the so-called Neyman orthogonality of φ makes it insensitive to inserting potentially biased machine learning estimators of the nuisance parameters; please see Lemma 2.E.1 in the appendix. A function is Neyman orthogonal if its Gateaux derivative, which is a directional derivative, vanishes

at the true η^0. In particular, Neyman orthogonality is a first-order property. The product relationship of the machine learning estimating errors described above is used to bound second-order terms.

Our proof of Theorem 2.2.5 uses techniques presented by Chernozhukov et al. (2018) and a version of Stein's method (Stein, 1972) that bounds the error of the normal approximation of a sum of random variables that exhibit a certain dependency structure. We use Theorem 3.6 from Ross (2011) that is defined on the dependency graph of network data where the dependency structure among the units should not be too dense. This is captured by Assumption 2.A.2 that restricts the maximal degree $d_{\max}$ in the dependency graph G_D on the unit-level data S_i, $i \in [N]$ to $d_{\max} = o(N^{1/4})$. That is, $d_{\max}$ may grow at a slower rate than $N^{1/4}$, which allows us to consider increasingly complex networks as the sample size increases.

2.2.5 | Consistent Variance Estimator

Under the assumptions of Theorem 2.2.5 and additional ones stated in Theorem 2.2.6 below, the asymptotic variance σ^2_∞ in Theorem 2.2.5 can be estimated consistently. The challenge is that the unit-level direct effects θ^0_i for $i \in [N]$, given in (2.2.2), are not all equal. This is because the unit-level data points S_i are typically not identically distributed. The difference in distributions originate from the X- and Z- features that generally depend on a varying number of other units. If two unit-level data points S_i and S_j have the same distribution, then their unit-level treatment effects θ^0_i and θ^0_j coincide. If enough of these unit-level treatment effects coincide, we can use the corresponding unit-level data to estimate them. Subsequently, we describe this procedure.

We partition $[N]$ into sets $\mathcal{A}_d$ for $d \geq 0$ such that all unit-level data points S_i for $i \in \mathcal{A}_d$ have the same distribution. Provided that the sets $\mathcal{A}_d$ are large enough, we can consistently estimate the corresponding θ^0_d for $d \geq 0$ by

$$\hat{\theta}_d = \frac{1}{|\mathcal{A}_d|} \sum_{i \in \mathcal{A}_d} \varphi(S_i, \hat{\eta}^{I^c_{k(i)}}), \tag{2.12}$$

where $k(i)$ denotes the index in $[K]$ such that $i \in I_{k(i)}$. The convergence rate of these estimators is at least $N^{-1/4}$; see Lemma 2.F.2 in Section 2.F in the appendix. To achieve this rate, we require that the sets $\mathcal{A}_d$ contain at least of order $N^{3/4}$ many indices; see Assumption 2.A.5 in Section 3.B in the appendix. The parametric convergence rate cannot be achieved in general because $\mathcal{A}_d$ is of smaller size than N, but the corresponding units may have the maximal $d_{\max}$ many ties in the network.

Subsequently, we characterize a situation in which the index d corresponds to the degree in the dependency graph G_D. This is the case if two unit-level data

points S_i and S_j have the same distribution if and only if the units i and j have the same degree in G_D. We assume, given a unit i and some $m \in [N] \setminus \{i\}$, that 1) if C_m is part of Z_i, then C_m is also part of X_i and vice versa; and 2) if W_m is part of X_i, then C_m is part of X_i and Z_i and vice versa. Consequently, if two units $i \neq j$ have the same degree in the dependency graph, then their X- and their Z-features are computed using the same number of random variables. Hence, X_i and X_j as well as Z_i and Z_j are identically distributed, and therefore S_i and S_j have the same distribution. Thus, the sets $\mathcal{A}_d$ form partition of the units according to their degree in the dependency graph, that is, $\mathcal{A}_d = \{i \in [N] \colon d(i) = d\}$ for $d \geq 0$, where $d(i)$ denotes the degree of i in the dependency graph. There are $d_{\max} + 1 = o(N^{1/4})$ many such sets, and each of them is required to be of size at least of order $N^{3/4}$ in Lemma 2.F.2. This is feasible because there are N units in total. Provided that the machine learning estimators of the nuisance functions converge at a rate faster than $N^{1/4}$ as specified by Assumption 2.A.6 in the appendix, we have the following consistent estimator of the asymptotic variance given in Theorem 2.2.6. Algorithm 3 summarizes the whole procedure of point estimation and inference for the EATE where the variance is estimated as given in Theorem 2.2.6. Nevertheless, this estimation scheme can be extended to general sets $\mathcal{A}_d$.

Theorem 2.2.6. *Denote by $G_D = (V, E_D)$ the dependency graph on S_i, $i \in [N]$. For a unit $i \in [N]$, denote by $d(i)$ its degree in G_D and by $k(i)$ the number in $[K]$ such that $S_i \in I_{k(i)}$. In addition to the assumptions made in Theorem 2.2.5, also assume that Assumption 2.A.5 and 2.A.6 stated in Section 3.B in the appendix hold. Based on φ defined in (2.5), we define the score function $\psi(S_i, \theta, \eta) = \varphi(S_i, \eta) - \theta$ for some general $\theta \in \mathbb{R}$ and the nuisance function triple $\eta = (g_1, g_0, h)$. Then,*

$$\frac{1}{N} \sum_{i=1}^{N} \psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) + \frac{2}{N} \sum_{\{i,j\} \in E_D} \psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}})$$

is a consistent estimator of the asymptotic variance σ^2_∞ in Theorem 2.2.5.

2.3 | Empirical Validation

We demonstrate our method in a simulation study and on a real-world dataset. In the simulation study, we validate the performance of our method on different network structures and compare it to two popular treatment effect estimators. Afterwards, we investigate the effect of studying time on exam performance in the Swiss StudentLife Study (Stadtfeld et al., 2019; Vörös et al., 2021) taking into account the effect of social ties.

Algorithm 1: Estimating the EATE from observational data on networks with spillover effects using plugin machine learning

Input : N unit-level observations $S_i = (W_i, C_i, X_i, Z_i, Y_i)$ from the model (4.3), network G, feature functions f_x^l, $l \in [r]$ and f_z^l, $l \in [t]$, corresponding dependency graph G_D, natural number K, natural number B, significance level $\alpha \in [0, 1]$, machine learning algorithms.

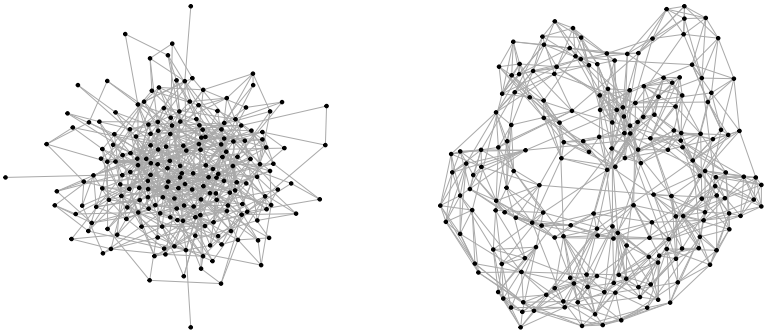
Output : Estimator of the EATE θ_N^0 and a valid p-value and confidence interval for the two-sided test $H_0 \colon \theta_N^0 = 0$ vs. $H_A \colon \theta_N^0 \neq 0$.

for $\underline{b \in [B]}$ **do**
  Randomly split the index set $[N]$ into K sets $I_1, \ldots, I_K$ of approximately equal size.
  for $\underline{k \in [K]}$ **do**
    Compute nuisance function estimators $\hat{g}_1^{I_k^c}$, $\hat{g}_0^{I_k^c}$, and $\hat{h}^{I_k^c}$ with machine learning algorithm and data from $\mathcal{S}_{I_k^c}$.
  end
  Compute point estimator of θ_N^0 according to (2.9), and call it $\hat{\theta}_b$.
  For degrees $d \geq 0$ in G_D, compute treatment effects $\hat{\theta}_d$ according to (2.12), and call them $\hat{\theta}_{d,b}$.
  Estimate asymptotic variance of $\hat{\theta}_b$ according to Theorem 2.2.6 using $\hat{\theta}_{d,b}$, and call it $\hat{\sigma}_{\infty,b}^2$.
  Compute p-value p_b for the two-sided test $H_0 \colon \theta_N^0 = 0$ vs. $H_A \colon \theta_N^0 \neq 0$ using $\hat{\theta}_b$, $\hat{\sigma}_{\infty,b}^2$, and asymptotic Gaussian approximation.
end
Compute $\hat{\theta} = \text{median}_{s \in [B]}(\hat{\theta}_b)$.
Compute aggregated p-value $p_{\text{aggr}}^0 = 2\,\text{median}_{b \in [B]}\, p_b$.
Compute confidence interval according to (2.10), call it $\text{CI}(\hat{\theta})$.
Return $\hat{\theta}$, p_{aggr}^0, $\text{CI}(\hat{\theta})$.

2.3.1 | Simulation Study

We compare the performance of our method to two popular alternative schemes with respect to bias of the point estimator and coverage and length of respective two-sided confidence intervals: the Hájek estimator and an IPW estimator. We first describe the two competitors and afterwards detail the simulation setting and present the results.

The **Hájek estimator** (denoted by "Hajek" in Figure 3.3.2) without incor-



(a) Erdős–Rényi network (b) Watts–Strogatz network

Figure 2.3: Different network structures on $N = 200$ units: Erdős–Rényi network where two nodes are connected with probability $6/N$ in 2.3a (every node is connected to 6 other nodes in expectation); Watts–Strogatz network with a rewiring probability of 0.05, a 1-dimensional ring-shaped starting lattice where each node is connected to 4 neighbors on both sides (that is, every node is connected to 8 other nodes), no loops, and no multiple edges in 2.3b. The graphs are generated using the R-package `igraph` (Csardi and Nepusz, 2006).

poration of confounders (Hájek, 1971) equals

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i Y_i}{\frac{1}{N}\Sigma_{j=1}^{N} W_i} + \frac{(1-W_i)Y_i}{\frac{1}{N}\Sigma_{j=1}^{N}(1-W_i)}\right).$$

The parametric convergence rate and asymptotic Gaussian distribution are preserved under X-spillover effects that equal the fraction of treated neighbors in a randomized experiment (Li and Wager, 2022). The **IPW estimator** (Rosenbaum, 1987) has been developed under SUTVA and uses observed confounding by creating a "pseudo population" in which the treatment is independent of the confounders (Hirano et al., 2003). We compute it using sample splitting and cross-fitting according to

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\left(\frac{W_i Y_i}{\hat{e}^{I_k^c}(C_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}^{I_k^c}(C_i)}\right),$$

where $\hat{e}^{I_k^c}$ is the fitted propensity score obtained by regressing W_i on C_i on the data in $i \in \mathcal{S}_{I_k^c}$. In our simulation, $\hat{e}^{I_k^c}$ coincides with $\hat{h}^{I_k^c}$ because we consider no Z-features. We denote this estimator by "IPW" in Figure 3.3.2

We investigate two network structures: Erdős–Rényi networks (Erdős and

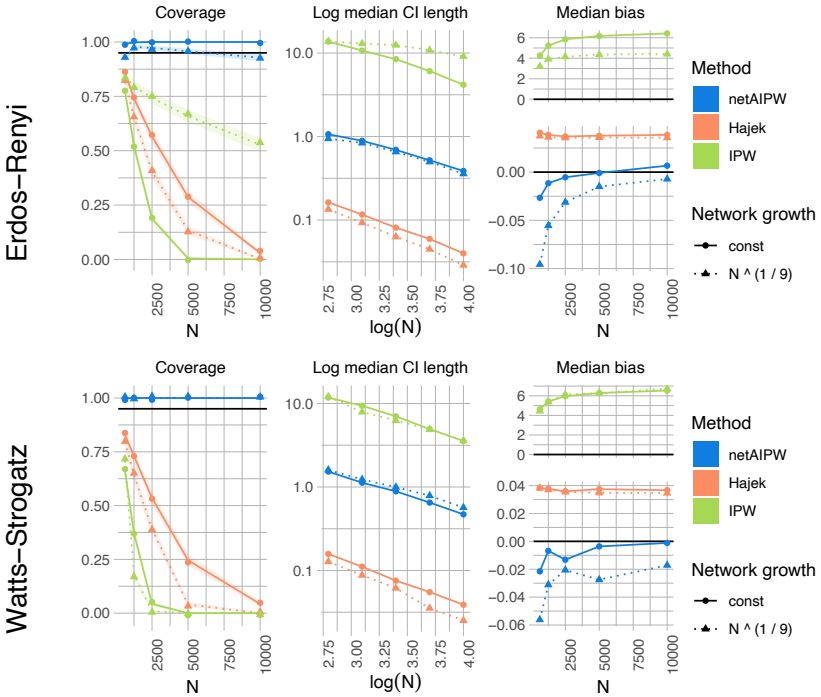


Figure 2.4: Coverage (fraction of times the true, and in general unknown, θ_N^0 was inside the confidence interval) and log median length of two-sided 95% confidence intervals for θ_N^0 and median bias over 1000 simulation runs for Erdős–Rényi and Watts–Strogatz networks of different complexities (Erdős–Rényi: expected degree 3 and $3N^{1/9}$ for "const" and "$N\hat{}(1/9)$", respectively; Watts–Strogatz: before rewiring, nodes have degree 4 and $4N^{1/9}$ for "const" and "$N\hat{}(1/9)$", respectively, and the rewiring probability is 0.05). We compare the performance of our method, netAIPW, with the Hájek and an IPW estimator, indicated by color. The variance of the competitors are empirical variances over the 1000 repetitions, whereas we computed confidence intervals for netAIPW according to (2.10). The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1000 simulation runs, and the dots in this panel are jittered (using `width = 0.02` and `height = 0.01`).

Rényi, 1959) and Watts–Strogatz networks (Watts and Strogatz, 1998). Erdős–Rényi networks randomly form edges between units with a fixed probability and are a simple example of a random mathematical network model. These

networks play an important role as a standard against which to compare more complicated models. Watts–Strogatz networks, also called small-world networks, share two properties with many networks in the real world: a small average shortest path length and a large clustering coefficient. To construct such a network, the vertices are first arranged in a regular fashion and linked to a fixed number of their neighbors. Then, some randomly chosen edges are rewired with a constant rewiring probability. A representative of each network type is provided in Figure 2.3. For each of these two network types, we consider one case where the dependency in the network does not increase with N (denoted by "const" in Figure 3.3.2) and one where it increases with N (denoted by $N\hat{}(1/9)$ in Figure 3.3.2).

The specific unit-level structural equations (4.3) we consider in this simulation study are specified in Section 2.C in the appendix. The functions g_1^0 and g_0^0 are step functions, and h^0 is a sigmoid function. We use the same model for all network types and complexities. We use a 1-dimensional X-feature but no Z-features. The feature X_i of unit i equals the average of the symmetrized confounders of its direct neighbors in G, denoted by $\alpha(i)$ (not containing i itself):

$$X_i = \frac{1}{|\alpha(i)|} \sum_{j \in \alpha(i)} (\mathbb{1}_{W_j=1} - \mathbb{1}_{W_j=0}) C_j,$$

if $\alpha(i)$ is non-empty, and 0 else.

For the sample sizes $N = 625, 1250, 2500, 5000, 10\,000$, we perform 1000 simulation runs redrawing the data according to the SEM, consider $B = 20$ and $K = 10$ in Algorithm 3, and estimate the nuisance functions by random forests consisting of 500 trees with a minimal node size of 5 and other default parameters using the R-package `ranger` (Wright and Ziegler, 2017). Our results for the Erdős–Rényi and Watts–Strogatz networks are displayed in Figure 3.3.2. Two different panels are used to display the results for different ranges of the bias of the methods. For all network types and complexities, we observe the following. The IPW estimator incurs a substantial bias. On the one hand, this IPW estimator does not account for network spillover. On the other hand, even under SUTVA, it is not Neyman orthogonal, which means we are not allowed to plug in machine learning estimators of nuisance functions, and it is known to have a poor finite-sample performance due to estimated propensity scores $\hat{e}^{I_k^c}$ that may be close to 0 or 1. The Hájek estimator incurs some bias because it does not adjust for observed confounding and assumes a randomized treatment instead. The bias of our method (denoted by "netAIPW" in Figure 3.3.2) decreases as the sample size increases. As the dependency graph becomes more complex, our method requires more observations to achieve a small bias because the data sets $\mathcal{S}_{I_k^c}$ in (2.8), which are used to estimate the nuisance functions,

are smaller in denser networks. In terms of coverage, the two competitors perform poorly, whereas our method guarantees coverage. The overcoverage of our method can be attributed to the conservative aggregation scheme of the p-values (Meinshausen et al., 2009).

2.3.2 | Empirical Analysis: Swiss StudentLife Study Data

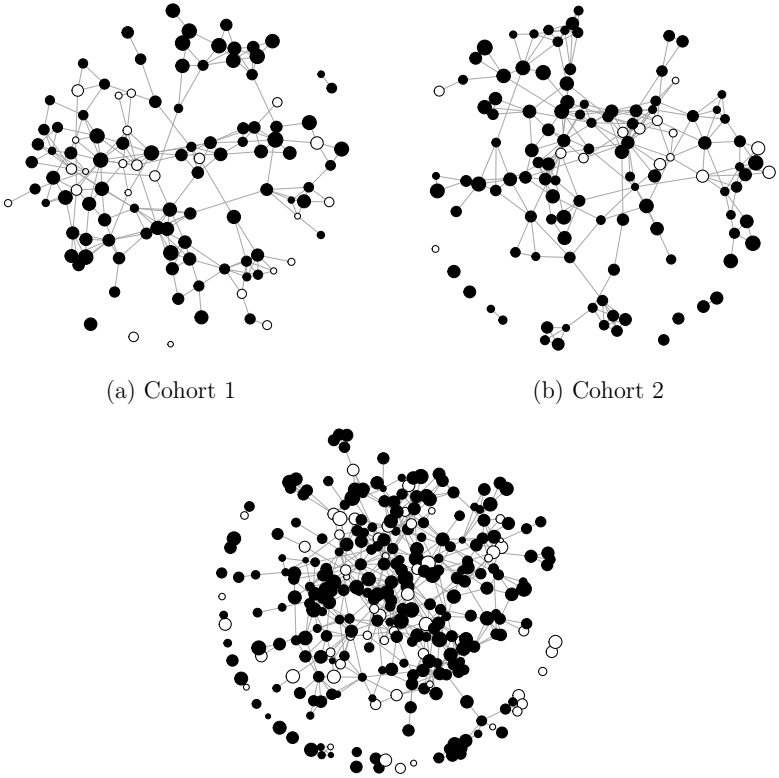
Subsequently, we estimate the causal effect of study time on academic success of university students with our newly developed estimator. We quantify this causal effect by the EATE that averages the difference in expected grade point average (GPA) of the final exam had a student studied much versus little, partialling out social network effects. Among the factors that determine academic success are person-specific traits, such as smartness (Chamorro-Premuzic and Furnham, 2008), willingness to work hard (Los and Schweinle, 2019), and the socioeconomic background (Heckman, 2006). Other factors are tied to the social embedding of a person (Stadtfeld et al., 2019). The Swiss StudentLife Study data (Stadtfeld et al., 2019; Vörös et al., 2021) was collected to study the impact of various factors on academic achievement. It consists of observations from freshmen undergraduate students pursuing a degree in the natural sciences at a Swiss university. Instead of a university entrance test, these students had to pass a demanding examination after one year of studying. At several timepoints throughout this year, the students were asked to fill out questionnaires about their student life, social network, and well-being. The data consists of three cohorts of students. Cohort 1 was observed in 2016 and cohorts 2 and 3 in 2017. Importantly, for all three cohorts, the data contains friendship information among the students. We build the corresponding undirected network by drawing an edge between two students if at least one of them mentioned the other one as being a friend. We believe that spillover effects arise due to students interacting in this network, and thus we have to control for them when estimating the EATE described above. Figure 2.5 displays the resulting network consisting of the three cohorts.

The GPA (Y_i) constitutes our response variable and represents the average grade of seven to nine exams, depending on study programs. It ranges from 1 to 6, with passing grades of 4 or higher. The average GPA in the data we used was 4.266 with a standard deviation of 0.872. The remaining variables were measured five to six months before the exam period and correspond to wave four of the Swiss StudentLife Study data. The self-reported number of hours spent studying per week during the semester (W_i) constitutes the treatment variable. It was dichotomized into studying many ($W_i = 1$) and few ($W_i = 0$) hours. We considered a setting where $W_i = 1$ corresponds to studying at least 8 hours per week, which is the 20% quantile, and one where $W_i = 1$ corresponds to studying at least 20 hours per week, which is the 80%

quantile. We consider spillover effects from the friends of a student, which are a student's direct neighbors in the friendship network. We consider Z-spillover effects that account for the effect of befriended students' study motivation and stress variables on a student's treatment. We do not consider spillover effects on the outcome GPA (no X-features). The Z_i-spillover variable of a student i is a vector of length 6, where each entry corresponds to the average of the following six variables across the friends of the student: (a) study motivation, measured with the learning objectives subscale of the SELLMO-ST[2] (Spinath et al., 2002), (b) work avoidance, measured with the work avoidance subscale of the students version of the SELLMO-ST[2], (c) the average of ten perceived stress items (Cohen and Williamson, 1988), (d, e) two items specifically on exam related stress, and (f) whether one was perceived as clever by at least one other student. In addition to these network effects, we control on the unit level (C_i) for the just mentioned variables observed on an individual unit as well as the cohort number, gender, having Swiss nationality, speaking German, and the financial situation. From all the data of the three cohorts combined, we only considered individuals for whom all the mentioned variables, that is, treatment, outcome, covariates, and Z-spillover variables, are observed. We did not perform missing value imputation. The final sample consisted of $N = 526$ individuals: 113 from cohort 1, 119 from cohort 2, and 294 from cohort 3. In our algorithm, we used $S = 1000$ sample splits with $K = 10$ groups each and random forests consisting of 5000 trees to learn g_0^0, g_1^0, and h^0 whose leaf size was initially determined by 5-fold cross-validation.

We estimated the EATE for different cutoffs in W_i of studying at least 8 and 20 hours per week, corresponding to the 20% and 80% quantiles, respectively. Table 2.1a displays our estimated EATE with $W_i = 1$ representing a weekly studying time of at least 8 hours. Our EATE estimator is positive and significant. On average, students received a 0.362 points higher GPA had they studied at least 8 hours per week compared to studying less. Consequently, a significantly higher GPA can be achieved by studying more. If we apply the same procedure but exclude the Z-spillover covariates (no spillover), the EATE estimator was higher and also significant. However, the higher effect estimator may be due to spurious association due to network spillover effects, highlighting the importance of controlling for such effects when estimating EATEs. Table 2.1b displays our results with $W_i = 1$ representing a weekly studying time of at least 20 hours. Our EATE estimator is positive but not significant anymore. Hence, our results suggest that the GPA is not significantly higher had a student studied at least 20 hours per week compared to studying less. Without spillover,

[2]This is a scale to assess learning and achievement motivation, and the subscale consists of eight items measured on a five-point Likert-scale from 1 ("completely disagree") to 5 ("completely agree").



(a) Cohort 1

(b) Cohort 2

(c) Cohort 3

Figure 2.5: Friendship networks per cohort with black dots representing $W_i = 1$ and a weekly studying time of at least 8 hours, white for $W_i = 0$ and a weekly studying time of less than 8 hours, and a bigger node size represents a higher GPA.

Spillover	EATE	95% CI for θ_N^0
yes	0.362	$[0.283, 0.442]$
no	0.451	$[0.364, 0.528]$

(a) $W_i = 1$ if studied at least 8 hours per week (20% quantile).

Spillover	EATE	95% CI for θ_N^0
yes	0.078	$[-0.096, 0.252]$
no	0.163	$[0.011, 0.311]$

(b) $W_i = 1$ if studied at least 20 hours per week (80% quantile).

Table 2.1: EATE and 95% confidence intervals for θ_N^0 for different settings with different control groups, namely studying less than 8 (a) or less than 20 (b) hours per week.

the treatment effect is significant. However, it is conceivable that spurious association due to network effects lead to this potentially biased result. Overall, the model including spillover effects seems more realistic than the one excluding them. Finally, when interpreting the results, it is important to recall that study time captures the learning time during the semester. There is an additional eight-week lecture-free preparation period, and our study time does not reflect this preparation time. Consequently, our results only describe the EATE of study time during the semester on GPA.

2.4 | Conclusion

Causal inference from observational data usually assumes independent units. However, having independent observations is often questionable, and so-called spillover effects among units are common in practice. Our aim was to develop point estimation and asymptotic inference for the expected average treatment effect (EATE) on observational network data. We would like to point out the hardness of this problem: we consider treatment effect estimation on data with increasing dependence among units, where the data generating mechanism can be highly nonlinear and include confounders. We use an augmented inverse probability weighting (AIPW) principle and account for spillover effects that we capture by features, which are functions of the known network and the treatment and covariate vectors.

Other authors who consider such a framework either uniformly limit the number of edges in the network, estimate densities, assume a semiparametric model, cannot incorporate observed confounding variables, assume the network consists of disconnected components, or limit interference to the direct neighbors in the network. Our AIPW machine learning approach overcomes these limitations. Units may interact beyond their direct neighborhoods, interactions may become increasingly complex as the sample size increases, and we consider arbitrary networks. We employ double machine learning techniques (Chernozhukov et al., 2018) to estimate the nuisance components of our model by arbitrary machine

learning algorithms. Although we employ machine learning algorithms, our EATE estimator converges at the $1/\sqrt{N}$-rate and asymptotically follows a Gaussian distribution, which allows us to perform inference.

In a simulation study, we demonstrated that commonly employed methods for treatment effect estimation suffer from the presence of spillover effects, whereas our method could account for the complex dependence structures in the data so that the bias vanished with increasing sample size and coverage was guaranteed. In the Swiss StudentLife Study, we investigated the EATE of study time on the grade point average of university examinations, accounting for spillover effects due to friendship relations. Omitting this spillover may lead to biased results due to spurious association.

In the present work, we focused on estimating the EATE. Other effects may be estimated in a similar manner, like for instance the global average treatment effect (GATE) where all units are jointly intervened on. Such an effect can compare giving the treatment to all units versus giving it to none. We develop an estimator of the GATE in Section 2.G in the appendix.

Acknowledgements

CE and PB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programm (grant agreement No. 786461), and M-LS received funding from the Swiss National Science Foundation (SNF) (project No. 200021_172485). The Swiss StudentLife data collection was supported by Swiss National Science Foundation Grant 10001A 169965 and the rectorate of ETH Zurich. We would like to thank Leonard Henckel and Dominik Rothenhäusler for constructive comments.

Appendix

2.A | Assumptions and Additional Definitions

We consider the following 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 $\xrightarrow{d}$ convergence in distribution. The symbol $\perp\!\!\!\perp$ denotes independence of random variables.

We observe N units according to the structural equations (4.3) that are connected by an underlying network. For each unit $i \in [N]$, we concatenate $S_i = (W_i, C_i, X_i, Z_i, Y_i)$ that are relevant for unit i. For notational simplicity, we abbreviate $D_i = (C_i, X_i)$ and $U_i = (C_i, Z_i)$ for $i \in [N]$.

Let the number of sample splits $K \geq 2$ be a fixed integer independent of N. We assume that $N \geq K$ holds. Consider a partition $I_1, \ldots, I_K$ of $[N]$. We assume that all sets $I_1, \ldots, I_K$ are of equal cardinality n. We make this assumption for the sake of notational simplicity, but our results hold without it.

Let $\{\delta_N\}_{N \geq K}$ and $\{\Delta_N\}_{N \geq K}$ be two sequences of non-negative numbers that converge to 0 as $N \to \infty$. Let $\{\mathcal{P}_N\}_{N \geq 1}$ be a sequence of sets of probability distributions P of the N units. We make the following additional sets of assumptions.

The following Assumption 3.B.2 recalls that we use the model (4.3) and specifies regularity assumptions on the involved random variables. Assumption 3.B.2.2 and 3.B.2.6 ensure that the random variables are integrable enough. Assumption 3.B.2.4 ensures that the true underlying function h^0 of the treatment assignment model is bounded away from 0 and 1, which allows us to divide by h^0 and $1 - h^0$.

Assumptions 2.A.1. *Let $p \geq 4$. For all N, all $i \in [N]$, all $P \in \mathcal{P}_N$, and all $k \in [K]$, we have the following.*

2.A.1.1 The structural equations (4.3) *hold, where the treatment $W_i \in \{0, 1\}$ is binary.*

2.A.1.2 There is a finite real constant C_1 independent of P satisfying $\|W_i\|_{P,p} + \|C_i\|_{P,p} + \|X_i\|_{P,p} + \|Z_i\|_{P,p} + \|Y_i\|_{P,p} \leq C_1$.

2.A.1.3 There is a finite real constant C_2 independent of P such that we have $\|Y_i\|_{P,\infty} + \|g_1^0(D_i)\|_{P,\infty} + \|g_0^0(D_i)\|_{P,\infty} + \|h^0(U_i)\|_{P,\infty} \leq C_2$.

2.A.1.4 There is a finite real constant C_3 independent of P such that $P(C_3 \leq h^0(U_i) \leq 1 - C_3) = 1$ holds.

2.A.1.5 There is a finite real constant C_4 such that we have $|\theta_i^0| \leq C_4$.

The following assumption limits the growth rate of the maximal degree of a node in the dependency graph.

Assumptions 2.A.2. *The maximal degree $d_{\max}$ of a node in the dependency graph satisfies $d_{\max} = o(N^{1/4})$. That is, $d_{\max}$ is allowed to grow at a slower rate than $N^{1/4}$ as $N \to \infty$.*

The following assumption allows us to characterize the asymptotic variance in Theorm 2.2.6 as the limit of the population variance on the N units.

Assumptions 2.A.3. *There is a finite real constant $\sigma_\infty^2 > 0$ such that for all $P \in \mathcal{P}_N$, we have*

$$\lim_{N\to\infty} Var\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0)\right) = \sigma_\infty^2.$$

The following Assumption 3.B.4 characterizes the realization set of the nuisance functions and the $N^{-1/2}$ convergence rate of products of the machine learning errors from estimating the nuisance functions g_1^0, g_0^0, and h^0.

Assumptions 2.A.4. *Consider the $p \geq 4$ from Assumption 3.B.2. For all $N \geq K$ and all $P \in \mathcal{P}_N$, consider a nuisance function realization set $\mathcal{T}$ such that the following conditions hold.*

2.A.4.1 The set $\mathcal{T}$ consists of P-integrable functions $\eta = (g_1, g_0, h)$ whose pth moment exists and whose $\|\cdot\|_{P,\infty}$-norm is uniformly bounded, and $\mathcal{T}$ contains $\eta^0 = (g_1^0, g_0^0, h^0)$. Furthermore, there is a finite real constant C_5 such that for all $i \in [N]$ and all elements $\eta = (g_0, g_1, h) \in \mathcal{T}$, we have

$$\|h^0(W_i) - h(W_i)\|_{P,2} \cdot \big(\|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \|g_0^0(D_i) - g_0(D_i)\|_{P,2} + \|h^0(W_i) - h(W_i)\|_{P,2}\big) \leq \delta_N N^{-\frac{1}{2}}.$$

2.A.4.2 Assumption 3.B.2.4 also holds with h^0 replaced by h.

2.A.4.3 Let κ be the largest real number such that for all $i \in [N]$ and all $\eta \in \mathcal{T}$, we have

$$\|h^0(W_i) - h(W_i)\|_{P,2} + \|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \|g_0^0(D_i) - g_0(D_i)\|_{P,2} \lesssim \sqrt{\delta_N} N^{-\kappa}.$$

That is, κ represents the slowest convergence rate of our machine learners. Then, there is a finite real constant C_6 exists such that $d_{\max} N^{-2\kappa} \leq C_6$ holds, where $d_{\max}$ denotes the maximal degree of the dependency graph.

2.A.4.4 For all $k \in [K]$, the nuisance parameter estimate $\hat{\eta}^{I_k} = \hat{\eta}^{I_k^c}(\mathcal{S}_{I_k^c})$ belongs to the nuisance function realization set $\mathcal{T}$ with P-probability no less than $1 - \Delta_N$.

The following two assumptions, Assumption 2.A.5 and 2.A.6, are only required to establish that our estimator of the asymptotic variance is consistent. They are not required to establish the asymptotic Gaussian distribution of our plugin machine learning estimator.

Assumption 2.A.5 characterizes the order of the minimal size of the sets $\mathcal{A}_d$ for $d \geq 0$. These sets are required to contain a sufficient number of units such that the degree-specific treatment effects θ_d^0 for $d \geq 0$ can be estimated at a fast enough rate. These estimators are required to give a consistent estimator of the asymptotic variance σ_∞^2.

Assumptions 2.A.5. *For $d \geq 0$, the order of $|\mathcal{A}_d|$ is at least $N^{3/4}$, denoted by $\Omega(N^{3/4})$ according to the Bachmann–Landau notation (Lattimore and Szepesvári, 2020).*

Assumption 2.A.6 specifies that all individual machine learning estimators of the nuisance functions converge at a rate faster than $N^{-1/4}$.

Assumptions 2.A.6. *The slowest convergence rate κ in Assumption 3.B.4.3 satisfies $\kappa \geq 1/4$.*

2.B | Network Effects in the Social Sciences

We consider models related to the term spillover effects. However, another notion of spillover effects has prevailed within the social science networks literature, namely social influence effects. In this appendix, we describe social influence effects and how their modeling differs from our approach. Whereas spillover effects represent new covariates on the unit-level that are built from variables of other units along network paths, social influence effects mostly concern effects that a specific variable A_j of neighboring units has on A_i of the ith unit. In the statistics literature, this is called contagion (Ugander et al., 2013; Eckles et al., 2017). In the social sciences, there are two important models to investigate social influence / contagion processes: the autologistic actor attribute model (ALAAM; Robins et al. (2001); Daraganova and Robins (2012)) and the stochastic actor-oriented model (SAOM; Snijders (2005); Snijders et al. (2010); Steglich et al. (2010)). Both models aim at estimating the degree to

which a variable A_i of a focal individual is associated with the values of its neighbors' values of A. Whereas ALAAMs only considers cross-sectional data, SAOMs additionally allow estimating longitudinal social influence effects.

In contrast, the spillover features that we consider summarize variables from neighboring units. They represent a new variable that is used for the treatment or outcome regression models. For example, in our empirical analysis, we consider the spillover effect of study motivation of unit i's neighbors on the learning hours of unit i. We do not consider spillover from the learning hours of unit i's neighbors on unit i's own learning hours (i.e. social influence / contagion). Instead, we model such associations of the individual units' learning hours by constructing features from other variables and units that act as observed confounders. Moreover, we are not interested in estimating the effect as such of, say, other units' study motivation on the learning hours of unit i. However, this is possible with ALAAMs and SAOMs. We are not interested in estimating spillover as such, but we consider spillover as a tool to control for spurious associations due to the network structure to estimate treatment effects.

2.C | Structural Equation Model for Simulation

For each unit $i \in [N]$, we sample independent and identically distributed confounders $C_i \sim \text{Unif}(0, 1)$ from the uniform distribution. The treatment assignments W_i are drawn from a Bernoulli distribution with success probability $p_i = \text{sigmoid}(C_i - 0.25)$, where $\text{sigmoid}(x) = 1/(1 + e^{-x})$ for $x \in \mathbb{R}$ denotes the sigmoid function. Let $\alpha(i)$ denote the neighbors of unit i in the network (without i itself). Then, we let the feature X_i denote the shifted average number of neighbors assigned to treatment weighted by their confounder, namely

$$X_i = \frac{1}{|\alpha(i)|} \sum_{j \in \alpha(i)} (\mathbb{1}_{W_j=1} - \mathbb{1}_{W_j=0}) C_j$$

if $\alpha(i)$ is non-empty, and 0 else. For real numbers x and c, we consider the functions

$$g_1^0(x, c) = 1.5\mathbb{1}_{x \geq 0.5, x < 0.7} + 4\mathbb{1}_{x \geq 0.7} + 2.5\mathbb{1}_{x < 0.5}$$

and

$$g_0^0(x, c) = 0.5\mathbb{1}_{x \geq 0.4, c \geq 0.2} - 0.75\mathbb{1}_{x \geq 0.4, c < 0.2} + 0.25\mathbb{1}_{x < 0.4, c \geq 0.2} - 0.5\mathbb{1}_{x < 0.4, c < 0.2}.$$

We consider error terms $\varepsilon_{Y_i} \sim \text{Unif}(-\sqrt{0.12}/2, \sqrt{0.12}/2)$ that are independent and identically distributed, and we consider the outcomes $Y_i = W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i}$.

2.D | Supplementary Lemmata

In this section, we prove two results on conditional independence relationships of the variables from our model. We argue for the directed acyclic graph (DAG) of our model (4.3) and use graphical criteria (Lauritzen, 1996; Pearl, 1998, 2009, 2010; Peters et al., 2017; Perković et al., 2018; Maathuis et al., 2019).

Lemma 2.D.1. *Let $i \in [N]$, and let $C_j \notin \mathrm{pa}(Y_i)$. Then, we have $Y_i \perp\!\!\!\perp C_j | \mathrm{pa}(Y_i)$.*

Proof of Lemma 2.D.1. The parents of Y_i are a valid adjustment set (Pearl, 2009). Because Y_i has no descendants, the claim follows. □

Lemma 2.D.2. *Let $i \in [N]$, and let $C_j \notin \mathrm{pa}(W_i)$. Then, we have $W_i \perp\!\!\!\perp C_j | \mathrm{pa}(W_i)$. Furthermore, for $j \neq i$, we have $W_i \perp\!\!\!\perp W_j | \mathrm{pa}(W_i)$.*

Proof of Lemma 2.D.2. The parents of W_i are a valid adjustment set (Pearl, 2009). The treatment variable W_i has no descendants apart from responses Y, which are colliders on any path from W_i to C_j or W_j, and thus the empty set blocks these paths. Consequently, the two claims follow. □

2.E | Proof of Theorem 2.2.5

Proof of Lemma 2.2.2. Let $i \in [N]$. We have

$$\mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)] = \mathbb{E}\left[\frac{W_i}{h^0(U_i)}\big(Y_i - g_1^0(D_i)\big)\right] - \mathbb{E}\left[\frac{1 - W_i}{1 - h^0(U_i)}\big(Y_i - g_0^0(D_i)\big)\right].$$

We have

$$\begin{aligned} & \mathbb{E}\left[\tfrac{W_i}{h^0(U_i)}\big(Y_i - g_1^0(D_i)\big)\right] \\ = \; & \mathbb{E}\left[\tfrac{W_i}{h^0(U_i)}\big(\mathbb{E}[Y_i | \mathrm{pa}(Y_i) \cup \mathrm{pa}(W_i)] - g_1^0(D_i)\big)\right] \\ = \; & \mathbb{E}\left[\tfrac{1}{h^0(U_i)}\mathbb{E}[W_i Y_i - W_i g_1^0(D_i) | \mathrm{pa}(Y_i)]\right] \\ = \; & \mathbb{E}\left[\tfrac{W_i}{h^0(U_i)}\mathbb{E}[\varepsilon_{Y_i} | \mathrm{pa}(Y_i)]\right] \\ = \; & 0 \end{aligned} \tag{2.13}$$

due to Lemma 2.D.1 and because $\mathbb{E}[\varepsilon_{Y_i} | \mathrm{pa}(Y_i)] = 0$ holds by assumption. Analogous computations for $\mathbb{E}[(1 - W_i)/(1 - h^0(U_i))(Y_i - g_0^0(D_i))]$ conclude the proof. □

The following lemma shows that the score function φ is Neyman orthogonal in the sense that its Gateaux derivative vanishes (Chernozhukov et al., 2018).

Lemma 2.E.1 (Neyman orthogonality)**.** *Assume the assumptions of Theorem 2.2.5 hold. Let $\eta \in \mathcal{T}$, and let $i \in [N]$. Then, we have*

$$\frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}\left[\varphi\big(S_i, \eta^0 + r(\eta - \eta^0)\big)\right] = 0.$$

Proof of Lemma 2.E.1. Let $r \in (0,1)$, let $i \in [N]$, and let $\eta \in \mathcal{T}$. Then, we have

$$\begin{aligned}
&\tfrac{\partial}{\partial r}\,\mathbb{E}\left[\varphi\big(S_i, \eta^0 + r(\eta - \eta^0)\big)\right] \\
=\;& \tfrac{\partial}{\partial r}\,\mathbb{E}\Bigg[g_1^0(D_i) - g_0^0(D_i) + r\big(g_1(D_i) - g_0(D_i) - g_1^0(D_i) + g_0^0(D_i)\big) \\
&\quad + \tfrac{W_i}{h^0(U_i) + r\left(h(U_i) - h^0(U_i)\right)}\Big(Y_i - g_1^0(D_i) - r\big(g_1(D_i) - g_1^0(D_i)\big)\Big) \\
&\quad - \tfrac{1 - W_i}{1 - h^0(U_i) - r\left(h(U_i) - h^0(U_i)\right)}\Big(Y_i - g_0^0(D_i) - r\big(g_0(D_i) - g_0^0(D_i)\big)\Big)\Bigg] \\
=\;& \mathbb{E}\Bigg[\big(g_1(D_i) - g_0(D_i)\big) - \big(g_1^0(D_i) - g_0^0(D_i)\big) \\
&\quad + \tfrac{W_i}{\left(h^0(U_i) + r\left(h(U_i) - h^0(U_i)\right)\right)^2}\Big(- \big(g_1(D_i) - g_1^0(D_i)\big) \\
&\qquad \cdot\big(h^0(U_i) + r(h(U_i) - h^0(U_i))\big) \\
&\qquad - \big(Y_i - g_1^0(D_i) - r(g_1(D_i) - g_1^0(D_i))\big)\big(h(U_i) - h^0(U_i)\big)\Big) \\
&\quad - \tfrac{1 - W_i}{\left(1 - h^0(U_i) - r\left(h(U_i) - h^0(U_i)\right)\right)^2}\Big(- \big(g_0(D_i) - g_0^0(D_i)\big) \\
&\qquad \cdot\big(1 - h^0(U_i) - r(h(U_i) - h^0(U_i))\big) \\
&\qquad + \big(Y_i - g_0^0(D_i) - r(g_0(D_i) - g_0^0(D_i))\big)\big(h(U_i) - h^0(U_i)\big)\Big)\Bigg].
\end{aligned} \tag{2.14}$$

We evaluate this expression at $r = 0$ and obtain

$$\begin{aligned}
&\tfrac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}\left[\varphi\big(S_i, \eta^0 + r(\eta - \eta^0)\big)\right] \\
=\;& \mathbb{E}\Bigg[\big(g_1(D_i) - g_0(D_i)\big) - \big(g_1^0(D_i) - g_0^0(D_i)\big) \\
&\quad - \Big(1 + \tfrac{\varepsilon_{W_i}}{h^0(U_i)}\Big)\big(g_1(D_i) - g_1^0(D_i)\big) - \tfrac{W_i}{(h^0(U_i))^2}\big(Y_i - g_1^0(D_i)\big)\big(h(U_i) - h^0(U_i)\big) \\
&\quad + \Big(1 - \tfrac{\varepsilon_{W_i}}{1 - h^0(U_i)}\Big)\big(g_0(D_i) - g_0^0(D_i)\big) \\
&\quad - \tfrac{1 - W_i}{(1 - h^0(U_i))^2}\big(Y_i - g_0^0(D_i)\big)\big(h(U_i) - h^0(U_i)\big)\Bigg] \\
=\;& 0
\end{aligned}$$

due to (2.13) and because

$$\begin{aligned}
&\mathbb{E}\left[\tfrac{\varepsilon_{W_i}}{h^0(U_i)}\big(g_1(D_i)-g_1^0(D_i)\big)\right]\\
=\;&\mathbb{E}\left[\big(\mathbb{E}[W_i|\operatorname{pa}(W_i)\cup\operatorname{pa}(Y_i)]-h^0(U_i)\big)\tfrac{1}{h^0(U_i)}\big(g_1(D_i)-g_1^0(D_i)\big)\right]\\
=\;&\mathbb{E}\left[\mathbb{E}[W_i-h^0(U_i)|\operatorname{pa}(W_i)]\tfrac{1}{h^0(U_i)}\big(g_1(D_i)-g_1^0(D_i)\big)\right]\\
=\;&\mathbb{E}\left[\mathbb{E}[\varepsilon_{W_i}|\operatorname{pa}(W_i)]\tfrac{1}{h^0(U_i)}\big(g_1(D_i)-g_1^0(D_i)\big)\right]\\
=\;&0
\end{aligned}$$

holds due to Lemma 2.D.2 and because we assumed $\mathbb{E}[\varepsilon_{W_i}|\operatorname{pa}(W_i)]=0$, and similarly for $\mathbb{E}[\varepsilon_{W_i}/(1-h^0(U_i))(g_0(D_i)-g_0^0(D_i))]$.

□

The following lemma bounds the second directional derivative of the score function. Its proof uses that products of the errors of the machine learners are of a smaller order than $N^{-1/2}$.

Lemma 2.E.2 (Product property)**.** *Assume the assumptions of Theorem 2.2.5 hold. Let $r\in(0,1)$, let $\eta\in\mathcal{T}$, and let $i\in[N]$. Then, we have*

$$\left|\frac{\partial^2}{\partial r^2}\,\mathbb{E}\left[\varphi\big(S_i,\eta^0+r(\eta-\eta^0)\big)\right]\right|\lesssim\delta_N N^{-\frac{1}{2}}.$$

Proof of Lemma 2.E.2. We use the first directional derivative we derived in (2.14) to compute the second directional derivative

$$\begin{aligned}
&\tfrac{\partial^2}{\partial r^2}\,\mathbb{E}\left[\varphi\big(S_i,\eta^0+r(\eta-\eta^0)\big)\right]\\
=\;&2\,\mathbb{E}\Bigg[\frac{W_i}{\Big(h^0(U_i)+r\big(h(U_i)-h^0(U_i)\big)\Big)^4}\Big(\big(g_1(D_i)-g_1^0(D_i)\big)\big(h^0(U_i)+r(h(U_i)-h^0(U_i))\big)\\
&\qquad+\big(Y_i-g_1^0(D_i)-r(g_1(D_i)-g_1^0(D_i))\big)\big(h(U_i)-h^0(U_i)\big)\Big)\\
&\qquad\cdot\Big(h^0(U_i)+r\big(h(U_i)-h^0(U_i)\big)\Big)\big(h(U_i)-h^0(U_i)\big)\Bigg]\\
&+2\,\mathbb{E}\Bigg[\frac{1-W_i}{\Big(1-h^0(U_i)-r\big(h(U_i)-h^0(U_i)\big)\Big)^4}\Big(\big(g_0(D_i)-g_0^0(D_i)\big)\\
&\qquad\cdot\big(1-h^0(U_i)-r(h(U_i)-h^0(U_i))\big)\\
&\qquad-\big(Y_i-g_0^0(D_i)-r(g_0(D_i)-g_0^0(D_i))\big)\big(h(U_i)-h^0(U_i)\big)\Big)\\
&\qquad\cdot\Big(1-h^0(U_i)-r\big(h(U_i)-h^0(U_i)\big)\Big)\big(h(U_i)-h^0(U_i)\big)\Bigg].
\end{aligned}$$

Due to Hölder's inequality and Assumption 3.B.2.1, 3.B.2.6, 3.B.2.4, and 3.B.4.1,

we have

$$\begin{aligned}
&\left|\tfrac{\partial^2}{\partial r^2}\,\mathbb{E}\left[\varphi\big(S_i,\eta^0+r(\eta-\eta^0)\big)\right]\right| \\
\lesssim\ &\big(\|g_1(D_i)-g_1^0(D_i)\|_{P,2}+\|h(U_i)-h^0(U_i)\|_{P,2}\big)\|h(U_i)-h^0(U_i)\|_{P,2} \\
&+\big(\|g_0(D_i)-g_0^0(D_i)\|_{P,2}+\|h(U_i)-h^0(U_i)\|_{P,2}\big)\|h(U_i)-h^0(U_i)\|_{P,2}.
\end{aligned}$$

Due to Assumption 3.B.4.1, both summands above are bounded by $\delta_N N^{-1/2}$, and hence we conclude the proof. □

The following lemma describes how we apply Stein's method (Chin, 2018) to obtain the asymptotic Gaussian distribution of our estimator although the data is highly dependent.

Lemma 2.E.3 (Asymptotic distribution with Stein's method)**.** *Assume the assumptions of Theorem 2.2.5 hold. Denote by*

$$\sigma_N^2 = Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi(S_i,\theta_i^0,\eta^0)\right).$$

Observe that by Assumption 2.A.3, we have $\lim_{N\to\infty}\sigma_N^2=\sigma_\infty^2>0$*. Then, we have*

$$\sigma_N^{-1}\cdot\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi(S_i,\theta_i^0,\eta^0)\xrightarrow{d}\mathcal{N}(0,1).$$

Proof of Lemma 2.E.3. According to Lemma 2.2.2, we have $\mathbb{E}[\psi(S_i,\theta_i^0,\eta^0)]=0$. According to Assumption 3.B.2, the fourth moment of $\psi(S_i,\theta_i^0,\eta^0)$ exists for all $i\in[N]$ and is uniformly bounded over $i\in[N]$. Recall that we denote by $d_{\max}$ the maximal degree in the dependency graph on S_i, $i\in[N]$. Due to Chin (2018, Lemma 1), we can thus bound the Wasserstein distance of $\sigma_N^{-1}\cdot\frac{1}{\sqrt{N}}\Sigma_{i=1}^N\psi(S_i,\theta_i^0,\eta^0)$ to $\mathcal{N}(0,1)$ as follows: there exist finite real constants c_1 and c_2 such that we have

$$\begin{aligned}
&d_{\mathcal{W}}\left(\sigma_N^{-1}\cdot\tfrac{1}{\sqrt{N}}\Sigma_{i=1}^N\psi(S_i,\theta_i^0,\eta^0)\right) \\
\le\ & c_1\cdot\tfrac{d_{\max}^{3/2}}{\sigma_N^2}\cdot\sqrt{\Sigma_{i=1}^N\mathbb{E}\left[\left(\tfrac{1}{\sqrt{N}}\psi(S_i,\theta_i^0,\eta^0)\right)^4\right]} \\
&+c_2\cdot\tfrac{d_{\max}^2}{\sigma_N^3}\cdot\Sigma_{i=1}^N\mathbb{E}\left[\left|\tfrac{1}{\sqrt{N}}\psi(S_i,\theta_i^0,\eta^0)\right|^3\right] \qquad (2.15)\\
=\ & c_1\cdot\tfrac{d_{\max}^{3/2}\cdot\frac{1}{\sqrt{N}}}{\sigma_N^2}\cdot\sqrt{\tfrac{1}{N}\Sigma_{i=1}^N\mathbb{E}[\psi^4(S_i,\theta_i^0,\eta^0)]} \\
&+c_2\cdot\tfrac{d_{\max}^2\cdot\frac{1}{\sqrt{N}}}{\sigma_N^{3/2}}\cdot\tfrac{1}{N}\Sigma_{i=1}^N\mathbb{E}\left[|\psi(S_i,\theta_i^0,\eta^0)|^3\right].
\end{aligned}$$

By assumption, we have $d_{\max}=o(N^{1/4})$. Thus, we have $d_{\max}^{3/2}\cdot\frac{1}{\sqrt{N}}=$

$o(N^{-1/8})$ and $d_{\max}^2 \cdot \frac{1}{\sqrt{N}} = o(1)$. Because the terms $\mathbb{E}[\psi^4(S_i, \theta_i^0, \eta^0)]$ and $\mathbb{E}\left[|\psi(S_i, \theta_i^0, \eta^0)|^3\right]$ are uniformly bounded over all $i \in [N]$ and because $\sigma_N \to \sigma_\infty$ as $N \to \infty$ according to Assumption 2.A.3, the Wasserstein distance in (2.15) is of order $o(1)$. Consequently, we infer the statement of the lemma. □

Lemma 2.E.4 (Vanishing covariance due to sparse dependency graph)**.** *Assume the assumptions of Theorem 2.2.5 hold. Let $k \in [K]$, and recall that $n = |I_k|$ holds. Then, we have*

$$\left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\varphi(S_i, \hat{\eta}^{I_k^c}) - \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}]\big) - \frac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\big) \right| = o_P(1).$$

Proof of Lemma 2.E.4. Let $k \in [K]$. We have

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\varphi(S_i, \hat{\eta}^{I_k^c}) - \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}]\big) \\
&\quad - \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\big)\Big|^2 \Big| \mathcal{S}_{I_k^c}\Big] \\
= \; & \tfrac{1}{n} \textstyle\sum_{i \in I_k} \mathbb{E}\left[\big(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\big)^2 | \mathcal{S}_{I_k^c}\right] - \tfrac{1}{n} \textstyle\sum_{i \in I_k} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}]^2 \\
&\quad + \tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}\left[\big(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\big)\big(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)\big) | \mathcal{S}_{I_k^c}\right] \\
&\quad - \tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}]\, \mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}].
\end{aligned} \tag{2.16}$$

Let $i \in [N]$. The nuisance parameter estimator $\hat{\eta}^{I_k^c}$ belongs to $\mathcal{T}$ with P-probability at least $1 - \Delta_N$ by Assumption 3.B.4.2. Therefore, with P-probability at least $1 - \Delta_N = 1 - o(1)$, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\big(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\big)^2 | \mathcal{S}_{I_k^c}\right]} \\
\leq \; & \sup_{\eta \in \mathcal{T}} \Big\| - g_1^0(D_i) + g_1(D_i) + g_0^0(D_i) - g_0(D_i) + \tfrac{W_i}{h^0(U_i)} \varepsilon_{Y_i} \\
&\quad - \tfrac{W_i}{h(U_i)} \big(g_1^0(D_i) - g_1(D_i) + \varepsilon_{Y_i}\big) - \tfrac{1 - W_i}{1 - h^0(U_i)} \varepsilon_{Y_i} \\
&\quad + \tfrac{1 - W_i}{1 - h(U_i)} \big(g_0^0(D_i) - g_0(D_i) + \varepsilon_{Y_i}\big) \Big\|_{P,2} \\
\leq \; & \sup_{\eta \in \mathcal{T}} \|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \sup_{\eta \in \mathcal{T}} \|g_0^0(D_i) - g_0(D_i)\|_{P,2} \\
&\quad + \sup_{\eta \in \mathcal{T}} \Big\| \tfrac{h(U_i) - h^0(U_i)}{h^0(U_i) h(U_i)} W_i \varepsilon_{Y_i} \Big\|_{P,2} + \sup_{\eta \in \mathcal{T}} \Big\| \tfrac{W_i}{h(U_i)} \big(g_1^0(D_i) - g_1(D_i)\big) \Big\|_{P,2} \\
&\quad + \sup_{\eta \in \mathcal{T}} \Big\| \tfrac{h^0(U_i) - h(U_i)}{(1 - h^0(U_i))(1 - h(U_i))} (1 - W_i) \varepsilon_{Y_i} \Big\|_{P,2} \\
&\quad + \sup_{\eta \in \mathcal{T}} \Big\| \tfrac{1 - W_i}{1 - h(U_i)} \big(g_0^0(D_i) - g_0(D_i)\big) \Big\|_{P,2}.
\end{aligned}$$

Assumption 3.B.2.1, 3.B.2.6, 3.B.2.4, and 2.A.4.2 bound the three terms $\|W_i \varepsilon_{Y_i} / (h^0(U_i) h(U_i))\|_{P,\infty}$, $\|W_i / h(U_i)\|_{P,\infty}$, $\|(1 - W_i)\varepsilon_{Y_i} / ((1 - h^0(U_i))(1 - h(U_i)))\|_{P,\infty}$, and $\|(1 - W_i)/(1 - h(U_i))\|_{P,\infty}$. Assumption 3.B.4.3 specifies that the error terms $\|h^0(W_i) - h(W_i)\|_{P,2}$, $\|g_1^0(D_i) - g_1(D_i)\|_{P,2}$, and $\|g_0^0(D_i) - g_0(D_i)\|_{P,2}$ are upper bounded by $\sqrt{\delta_N} N^{-\kappa}$. Due to Hölder's in-

equality, we infer

$$\sqrt{\mathbb{E}\left[\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\right)^2 \middle| \mathcal{S}_{I_k^c}\right]} \lesssim \sqrt{\delta_N} N^{-\kappa} \tag{2.17}$$

with P-probability at least $1 - \Delta_N$.

Subsequently, we bound the summands in (2.16). Due to (2.17), we have

$$\begin{aligned} &\tfrac{1}{n} \textstyle\sum_{i \in I_k} \mathbb{E}\left[\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\right)^2 \middle| \mathcal{S}_{I_k^c}\right] - \tfrac{1}{n} \textstyle\sum_{i \in I_k} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}]^2 \\ &\lesssim \delta_N N^{-2\kappa} \end{aligned}$$

with P-probability at least $1 - \Delta_N$. Observe that we have

$$\begin{aligned} &\tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}\left[\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\right)\left(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)\right) \middle| \mathcal{S}_{I_k^c}\right] \\ &\quad - \tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}] \, \mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}] \\ = \; &\tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \operatorname{Cov}\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) \middle| \mathcal{S}_{I_k^c}\right) \\ = \; &\tfrac{1}{n} \textstyle\sum_{i,j \in I_k, \{i,j\} \in E_D} \operatorname{Cov}\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) \middle| \mathcal{S}_{I_k^c}\right), \end{aligned}$$

where E_D denotes the edge set of the dependency graph, because the S_i with $i \in I_k$ are independent of data in $\mathcal{S}_{I_k^c}$ and because, given $\mathcal{S}_{I_k^c}$, $\varphi(S_i \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)$ and $\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)$ are uncorrelated if there is no edge between i and j in the dependency graph. In the dependency graph, each node has a maximal degree of $d_{\max}$. Thus, there are at most $1/2 \cdot N \cdot d_{\max}$ many edges in E_D. With P-probability at least $1 - \Delta_N$, the term

$$\operatorname{Cov}\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) \middle| \mathcal{S}_{I_k^c}\right)$$

can be bounded by $\delta_N N^{-2\kappa}$ up to constants for all i and j due to (2.17). Therefore, with P-probability at least $1 - \Delta_N$, we have

$$\begin{aligned} &\tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}\left[\left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)\right)\left(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)\right) \middle| \mathcal{S}_{I_k^c}\right] \\ &\quad - \tfrac{1}{n} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}] \, \mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}] \\ \lesssim \; &\delta_N d_{\max} N^{-2\kappa} \\ \lesssim \; &\delta_N, \end{aligned}$$

where the last bound holds due to Assumption 3.B.4.3. Consequently, we have

$$\begin{aligned} \mathbb{E}\Bigg[\Bigg| &\tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \left(\varphi(S_i, \hat{\eta}^{I_k^c}) - \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}]\right) \\ &- \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \left(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\right)\Bigg|^2 \Bigg| \mathcal{S}_{I_k^c}\Bigg] \lesssim \delta_N \end{aligned}$$

with P-probability at least $1 - \Delta_N$, and we infer the statement of the lemma due to Chernozhukov et al. (2018, Lemma 6.1). □

Lemma 2.E.5 (Taylor expansion)**.** *Assume the assumptions of Theorem 2.2.5 hold. Let $k \in [K]$. We have*

$$\left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)] \big) \right| = o_P(1)$$

Proof of Lemma 2.E.5. Let $k \in [K]$. For $r \in [0,1]$, let us define the function

$$f_k(r) = \frac{1}{n} \sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \eta^0 + r(\hat{\eta}^{I_k^c} - \eta^0)) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)].$$

We have

$$\begin{aligned} & \mathbb{E}\left[\left| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)] \big) \right| \middle| \mathcal{S}_{I_k^c} \right] \\ = & \left| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)] \big) \right| \\ = & \sqrt{n} |f_k(1)|. \end{aligned}$$

We apply a Taylor expansion to $f_k(1)$ at 0 and obtain

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

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

$$\begin{aligned} & \mathbb{E}\left[\left| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)] \big) \right| \middle| \mathcal{S}_{I_k^c} \right] \\ \leq & \sqrt{n} \Big(|f_k(0)| + |f_k'(0)| + \sup_{r \in (0,1)} \tfrac{1}{2} |f_k''(r)| \Big). \end{aligned}$$

Due to the definition of f_k, we have $f_k(0) = 0$. Due to Neyman orthogonality that we established in Lemma 2.E.1, we have $f_k'(0) = 0$. Due to the product property that we established in Lemma 2.E.2, we have $\sup_{r \in (0,1)} \frac{1}{2} |f_k''(r)| \lesssim \delta_N N^{-1/2}$ with P-probability at least $1 - \Delta_N$ because $\hat{\eta}^{I_k^c}$ belongs to $\mathcal{T}$ with P-probability at least $1 - \Delta_N$. Consequently, we have

$$\mathbb{E}\left[\left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\varphi(S_i, \eta^0)] \big) \right| \middle| \mathcal{S}_{I_k^c} \right] \lesssim \delta_N$$

with P-probability at least $1 - \Delta_N$. We infer the statement of the lemma due to Chernozhukov et al. (2018, Lemma 6.1). □

Proof of Theorem 2.2.5. We have

$$\begin{aligned}
&\sqrt{N}(\hat{\theta} - \theta_N^0) \\
=\ & \sqrt{N} \cdot \tfrac{1}{nK} \textstyle\sum_{k=1}^K \sum_{i \in I_k} \psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) \\
=\ & \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \tfrac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)\big) + \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N \psi(S_i, \theta_i^0, \eta^0)
\end{aligned}$$

because the disjoint sets I_k are of equal size n, so that we have $N = nK$. Let $k \in [K]$. We have

$$\begin{aligned}
&\left| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)\big) \right| \\
\leq\ & \Big| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \mathbb{E}[\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}]\big) \\
&\qquad - \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\psi(S_i, \theta_i^0, \eta^0) - \mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)]\big) \Big| \\
&\quad + \Big| \tfrac{1}{\sqrt{n}} \textstyle\sum_{i \in I_k} \big(\mathbb{E}[\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)]\big) \Big| \\
=\ & o_P(1)
\end{aligned}$$

due to Hölder's inequality and Lemma 2.E.4 and 2.E.5. Because K is a constant independent of N, we have

$$\frac{1}{\sqrt{K}} \sum_{k=1}^K \frac{1}{\sqrt{n}} \sum_{i \in I_k} \big(\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)\big) = o_P(1).$$

Due to Lemma 2.E.3, we have $\frac{1}{\sqrt{N} \cdot \sigma_N} \sum_{i=1}^N \psi(S_i, \theta_i^0, \eta^0) \xrightarrow{d} \mathcal{N}(0, 1)$ as $N \to \infty$. We have $\sigma_N \to \sigma_\infty$ as $N \to \infty$ due to Assumption 2.A.3. Therefore, we have

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \psi(S_i, \theta_i^0, \eta^0) = \frac{1}{\sqrt{N} \cdot \sigma_N} \sum_{i=1}^N \psi(S_i, \theta_i^0, \eta^0) \cdot \sigma_N \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2)$$

as $N \to \infty$. Consequently, we have $\sqrt{N}(\hat{\theta} - \theta_N^0) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2)$ as claimed. □

2.F | Proof of Theorem 2.2.6

Lemma 2.F.1. *Assume the assumptions of Theorem 2.2.6 hold. Let $i \in [N]$. There exists a finite real constant C_7 independent of i such that $\|\psi(S_i, \theta_{d(i)}^0, \eta^0)\|_{P,4} \leq C_7$ holds. Consequently, for $i, j, m, r \in [N]$, we can also bound the following terms by finite uniform constants:*

- $\|\psi(S_i, \theta_{d(i)}^0, \eta^0)\|_{P,2}$
- $Var\big(\varphi(S_i, \eta^0)\big)$
- $Var\big(\psi^2(S_i, \theta_{d(i)}^0, \eta^0)\big)$
- $\mathrm{Cov}\big(\varphi(S_i, \eta^0), \varphi(S_j, \eta^0)\big)$

- $Var\big(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0)\big)$
- $\mathrm{Cov}\big(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0), \psi(S_m, \theta^0_{d(m)}, \eta^0)\psi(S_r, \theta^0_{d(r)}, \eta^0)\big)$

Moreover, we have $\varphi^2(S_i, \eta^0) = O_P(1)$ *and* $\psi^2(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) = O_P(1)$.

Proof of Lemma 2.F.1. We have

$$\begin{aligned} & \|\psi(S_i, \theta^0_{d(i)}, \eta^0)\|_{P,4} \\ \leq\ & \|g^0_1(D_i)\|_{P,4} + \|g^0_0(D_i)\|_{P,4} + \left\|\tfrac{W_i}{h^0(U_i)}\right\|_{P,4} \|Y_i - g^0_1(D_i)\|_{P,\infty} \\ & + \left\|\tfrac{1-W_i}{1-h^0(U_i)}\right\|_{P,4} \|Y_i - g^0_0(D_i)\|_{P,\infty} + |\theta^0_{d(i)}|. \end{aligned} \tag{2.18}$$

All individual summands in the above decomposition are bounded by a finite real constant independent of i due to Assumption 3.B.2. Therefore, there exists a finite real constant C_7 independent of i such that $\|\psi(S_i, \theta^0_i, \eta^0)\|_{P,4} \leq C_7$ holds.

The other terms in the statement of the present lemma are bounded as well by finite real constants independent of $i, j, m, r \in [N]$ due to Hölder's inequality.

Moreover, we have $\psi^2(S_i, \eta^0) = O_P(1)$ because $\|\psi^2(S_i, \eta^0)\|_{P,2}$ is bounded by a constant that is independent of i.

Furthermore, with P-probability at least $1 - \Delta_N$, we have

$$\mathbb{E}\left[\psi^2(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \middle| \mathcal{S}_{I^c_{k(i)}}\right] \leq \sup_{\eta \in \mathcal{T}} \mathbb{E}\left[\psi^2(S_i, \theta^0_{d(i)}, \eta)\right] = \sup_{\eta \in \mathcal{T}} \|\psi(S_i, \theta^0_{d(i)}, \eta)\|^2_{P,2}.$$

The term $\|\psi(S_i, \theta^0_{d(i)}, \eta)\|^2_{P,2}$ is bounded by a real constant that is independent of i and η because the derivation in (2.18) also holds with η^0 replaced by $\eta \in \mathcal{T}$ due to Assumption 3.B.4. □

Lemma 2.F.2 (Convergence rate of unit-level effect estimators)**.** *Assume the assumptions of Theorem 2.2.6 hold. Let* $d \geq 0$*, and assume that all assumptions of Section 3.B in the appendix hold. Then, we have* $\hat{\theta}_d - \theta^0_d = o_P(N^{-1/4})$*, where* $\hat{\theta}_d$ *is as in* (2.12).

Proof of Lemma 2.F.2. Let $d \geq 0$. Due to the definition of $\hat{\theta}_d$ given in (2.12) and Lemma 2.2.2, we have

$$\begin{aligned} & N^{\frac{1}{4}}(\hat{\theta}_d - \theta^0_d) \\ =\ & \tfrac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \textstyle\sum_{i \in \mathcal{A}_d} \big(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\varphi(S_i, \eta^0)]\big) \\ =\ & \tfrac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \textstyle\sum_{i \in \mathcal{A}_d} \big(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)\big) + \tfrac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \textstyle\sum_{i \in \mathcal{A}_d} \big(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\big). \end{aligned} \tag{2.19}$$

Subsequently, we show that the two sets of summands in (2.19) are of order $o_P(1)$. We start with the first set of summands. Let $i \in \mathcal{A}_d$. With P-probability

at least $1 - \Delta_N$, we have

$$\sqrt{\mathbb{E}\left[\left(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)\right)^2 | \mathcal{S}_{I^c_k}\right]} \lesssim \sqrt{\delta_N} N^{-\kappa}$$

due to Equation (2.17). Hence, we have $|\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| = O_P(\sqrt{\delta_N} N^{-\kappa})$ due to Chernozhukov et al. (2018, Lemma 6.1). Consequently, we have

$$\frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i \in \mathcal{A}_d} |\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| = O_P(\sqrt{\delta_N} N^{\frac{1}{4} - \kappa}) = o_P(1)$$

because we have $\kappa \geq 1/4$ by Assumption 2.A.6. Next, we show that the second set of summands in (2.19) is of order $o_P(1)$. Let $\varepsilon > 0$. We have

$$\begin{aligned}
& P\left(\left|\tfrac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \Sigma_{i \in \mathcal{A}_d} \left(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\right)\right|^2 > \varepsilon^2\right) \\
\leq & \tfrac{N^{\frac{1}{2}}}{\varepsilon^2 |\mathcal{A}_d|^2} \Big(\Sigma_{i \in \mathcal{A}_d} \operatorname{Var}(\varphi(S_i, \eta^0)) + \Sigma_{i,j \in \mathcal{A}_d, i \neq j} \operatorname{Cov}\left(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0)\right) \Big) \\
= & \tfrac{N^{\frac{1}{2}}}{\varepsilon^2 |\mathcal{A}_d|^2} (|\mathcal{A}_d| + 2|E_D \cap \mathcal{A}_d^2|) O(1)
\end{aligned}$$

because $\operatorname{Var}(\varphi(S_i, \eta^0))$ and $\operatorname{Cov}(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0))$ are bounded by constants uniformly over i due to Lemma 2.F.1, and because $\operatorname{Cov}(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0))$ does not equal 0 only if $\{i, j\} \in E_D \cap \mathcal{A}_d^2$, where E_D denotes the edge set of the dependency graph. There are $|\mathcal{A}_d|$ many nodes in $\mathcal{A}_d$, and each node has a maximal degree of $d_{\max}$. Thus, we have $|E_D \cap \mathcal{A}_d^2| \leq 1/2 |\mathcal{A}_d| d_{\max}$. Due to $d_{\max} = o(N^{1/4})$ and $|\mathcal{A}_d| = \Omega(N^{3/4})$, which hold according to Assumption 2.A.2 and 2.A.5, we obtain

$$\frac{N^{\frac{1}{2}}}{\varepsilon^2 |\mathcal{A}_d|^2} (|\mathcal{A}_d| + 2|E \cap \mathcal{A}_d^2|) O(1) = o(1).$$

Consequently, we also have

$$\left| \frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i \in \mathcal{A}_d} \left(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]\right) \right| = o_P(1).$$

□

Lemma 2.F.3 (Consistent variance estimator part I)**.** *Assume the assumptions of Theorem 2.2.6 hold. We have*

$$\left| \frac{1}{N} \sum_{i=1}^{N} \left(\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)]\right) \right| = o_P(1).$$

Proof of Lemma 2.F.3. We have

$$\begin{aligned} & \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\hat{\theta}_{d(i)},\hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\psi^2(S_i,\theta^0_{d(i)},\eta^0)]\big) \\ = \; & \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\hat{\theta}_{d(i)},\hat{\eta}^{I^c_{k(i)}}) - \psi^2(S_i,\hat{\theta}_{d(i)},\eta^0)\big) \\ & + \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\hat{\theta}_{d(i)},\eta^0) - \psi^2(S_i,\theta^0_{d(i)},\eta^0)\big) \\ & + \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\theta^0_{d(i)},\eta^0) - \mathbb{E}[\psi^2(S_i,\theta^0_{d(i)},\eta^0)]\big). \end{aligned} \tag{2.20}$$

We bound the three sets of summands in (2.20) individually. The first set of summands can be expressed as

$$\begin{aligned} & \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\hat{\theta}_{d(i)},\hat{\eta}^{I^c_{k(i)}}) - \psi^2(S_i,\hat{\theta}_{d(i)},\eta^0)\big) \\ = \; & \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\varphi^2(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i,\eta^0)\big) - \tfrac{2}{N}\Sigma_{i=1}^{N}\hat{\theta}_{d(i)}\big(\varphi(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i,\eta^0)\big). \end{aligned}$$

We have

$$\left|\frac{1}{N}\sum_{i=1}^{N}\big(\varphi^2(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i,\eta^0)\big)\right| = o_P(1) \tag{2.21}$$

because the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is continuous and due to Equation (2.17). Indeed, let $\varepsilon > 0$. Because the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is continuous, there exists $\delta > 0$ such that if $|\varphi(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i,\eta^0)| < \delta$, then also $|\varphi^2(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i,\eta^0)| < \varepsilon$. Consequently, we have

$$\begin{aligned} & P\big(|\varphi^2(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i,\eta^0)| > \varepsilon \big| \mathcal{S}_{I^c_{k(i)}}\big) \\ \leq \; & P\big(|\varphi(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i,\eta^0)| > \delta \big| \mathcal{S}_{I^c_{k(i)}}\big) \\ \leq \; & \tfrac{1}{\delta}\sup_{\eta\in\mathcal{T}}\|\varphi(S_i,\eta) - \varphi(S_i,\eta^0)\|_{P,1} \end{aligned}$$

with P-probability at least $1 - \Delta_N$, and we infer (2.21) due to (2.17). The estimator $\hat{\theta}_{d(i)}$ is a consistent estimator of $\theta^0_{d(i)}$ due to Lemma 2.F.2, and $\theta^0_{d(i)}$ is bounded independent of i due to Assumption 3.B.2.5. Moreover, we have $|\varphi(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i,\eta^0)| = o_P(1)$ due to (2.17) and Chernozhukov et al. (2018, Lemma 6.1). Consequently, we have

$$\left|\frac{2}{N}\sum_{i=1}^{N}\hat{\theta}_{d(i)}\big(\varphi(S_i,\hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i,\eta^0)\big)\right| = o_P(1)$$

due to Hölder's inequality. Hence, the first set of summands in (2.20) is of order $o_P(1)$. The second set of summand in (2.20) can be decomposed as

$$\begin{aligned} & \tfrac{1}{N}\Sigma_{i=1}^{N}\big(\psi^2(S_i,\hat{\theta}_{d(i)},\eta^0) - \psi^2(S_i,\theta^0_{d(i)},\eta^0)\big) \\ = \; & \tfrac{1}{N}\Sigma_{i=1}^{N}(\hat{\theta}^2_{d(i)} - (\theta^0_{d(i)})^2) - \tfrac{2}{N}\Sigma_{i=1}^{N}(\hat{\theta}_{d(i)} - \theta^0_{d(i)})\varphi(S_i,\eta^0). \end{aligned}$$

We have $|\frac{1}{N}\Sigma_{i=1}^{N}(\hat{\theta}^2_{d(i)} - (\theta^0_{d(i)})^2)| = o_P(1)$ due to Lemma 2.F.2. Lemma 2.F.1

bounds $\varphi^2(S_i, \eta^0)$ in probability. Due to Hölder's inequality, we obtain

$$\left| \frac{2}{N} \sum_{i=1}^{N} (\hat{\theta}_{d(i)} - \theta^0_{d(i)}) \varphi(S_i, \eta^0) \right| = o_P(1).$$

Consequently, the second set of summands in (2.20) is of order $o_P(1)$. Last, we bound the third set of summands in (2.20). Let $\varepsilon > 0$. We have

$$\begin{aligned}
& P\Big(\Big| \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \left(\psi^2(S_i, \theta^0_{d(i)}, \eta^0) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)] \right) \Big|^2 > \varepsilon^2 \Big) \\
\leq \; & \tfrac{1}{\varepsilon^2 N^2} \Big(\textstyle\sum_{i=1}^{N} \operatorname{Var}\left(\psi^2(S_i, \theta^0_{d(i)}, \eta^0) \right) \\
& \quad + \textstyle\sum_{i,j \in [N], \{i,j\} \in E_D} \operatorname{Cov}\left(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0) \right) \Big) \\
\leq \; & \tfrac{1}{\varepsilon^2 N^2} (NO(1) + N d_{\max} O(1)) \\
= \; & o(1)
\end{aligned}$$

because $\operatorname{Var}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0))$ and $\operatorname{Cov}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0))$ are bounded uniformly over i and j by Lemma 2.F.1, because we have that $\operatorname{Cov}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0))$ does not vanish only if $\{i, j\} \in E_D$, and because $d_{\max} = o(N^{1/4})$ by Assumption 2.A.2. Consequently, also the third set of summands in (2.20) is of order $o_P(1)$, and we have established the statement of the present lemma. □

Lemma 2.F.4 (Consistent variance estimator part II)**.** *Assume the assumptions of Theorem 2.2.6 hold. Denote by E_D the edge set of the dependency graph. We have*

$$\begin{aligned}
& \Big| \tfrac{1}{N} \textstyle\sum_{i,j \in [N], \{i,j\} \in E_D} \left(\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right. \\
& \qquad \left. - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)] \right) \Big| = o_P(1).
\end{aligned}$$

Proof of Lemma 2.F.4. We have the decomposition

$$\begin{aligned}
& \tfrac{1}{N} \textstyle\sum_{i,j \in [N], \{i,j\} \in E_D} \left(\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right. \\
& \quad \left. - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)] \right) \\
= \; & \tfrac{2}{N} \textstyle\sum_{\{i,j\} \in E_D} \left(\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right. \\
& \quad \left. - \psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right) \\
& + \tfrac{2}{N} \textstyle\sum_{\{i,j\} \in E_D} \left(\psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right. \\
& \quad \left. - \psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0) \right) \\
& + \tfrac{2}{N} \textstyle\sum_{\{i,j\} \in E_D} \left(\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0) \right. \\
& \quad \left. - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)] \right).
\end{aligned} \tag{2.22}$$

Subsequently, we bound the three sets of summands in (2.22) individually. We start by bounding the first set of summands. We have

$$\begin{aligned} & \tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}\big(\psi(S_i,\hat{\theta}_{d(i)},\hat{\eta}^{I^c_{k(i)}})\psi(S_j,\hat{\theta}_{d(j)},\hat{\eta}^{I^c_{k(j)}}) \\ & \quad -\psi(S_i,\theta^0_{d(i)},\hat{\eta}^{I^c_{k(i)}})\psi(S_j,\theta^0_{d(j)},\hat{\eta}^{I^c_{k(j)}})\big) \\ = \; & \tfrac{2}{N}\textstyle\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)}-\hat{\theta}_{d(i)})\psi(S_j,\theta^0_{d(j)},\hat{\eta}^{I^c_{k(j)}}) \\ & \quad +\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)}-\hat{\theta}_{d(i)})(\theta^0_{d(j)}-\hat{\theta}_{d(j)}). \end{aligned}$$

We have

$$\begin{aligned} & \left|\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)}-\hat{\theta}_{d(i)})\psi(S_j,\theta^0_{d(j)},\hat{\eta}^{I^c_{k(j)}})\right| \\ \leq \; & \sqrt{\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)}-\hat{\theta}_{d(i)})^2}\sqrt{\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}\psi(S_j,\theta^0_{d(j)},\hat{\eta}^{I^c_{k(j)}})} \\ = \; & \tfrac{1}{N}|E_D|o_P(N^{-1/4}) \\ = \; & d_{\max}o_P(N^{-1/4}) \\ = \; & o_P(1) \end{aligned}$$

due to Hölder's inequality, Lemma 2.F.2, Lemma 2.F.1, and Assumption 2.A.2. Moreover, we have

$$\left|\frac{1}{N}\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)}-\hat{\theta}_{d(i)})(\theta^0_{d(j)}-\hat{\theta}_{d(j)})\right| = \frac{1}{N}|E_D|o_P(N^{-1/2}) = o_P(1)$$

due to Hölder's inequality, Lemma 2.F.2, and Assumption 2.A.2. Consequently, the first set of summands in (2.22) is of order $o_P(1)$. We proceed to bound the second set of summands in (2.22). Let $\{i,j\}\in E_D$. Due to the construction of $\mathcal{S}_{I_{k(i)}}$ and $\mathcal{S}_{I^c_{k(i)}}$, we have $S_i=(W_i,C_i,X_i,Z_i,Y_i)\in\mathcal{S}_{I_{k(i)}}$, and none of W_i, C_i, Y_i, or the variables used to compute X_i belong to $\mathcal{S}_{I^c_{k(i)}}$. Moreover, the variables W_i, C_i, Y_i, and the variables used to compute X_i also cannot belong to $\mathcal{S}_{I^c_{k(j)}}$ as otherwise we would have $S_i \perp\!\!\!\perp S_j$, and consequently $\{i,j\}\notin E_D$. Therefore, we have

$$\begin{aligned} & \mathbb{E}\big[|\psi(S_i,\theta^0_{d(i)},\hat{\eta}^{I^c_{k(i)}})\psi(S_j,\theta^0_{d(j)},\hat{\eta}^{I^c_{k(j)}}) \\ & \quad -\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0)|\big|\mathcal{S}_{I^c_{k(i)}},\mathcal{S}_{I^c_{k(j)}}\big] \\ \leq \; & \sup_{\eta_1,\eta_2\in\mathcal{T}}\mathbb{E}\big[|\psi(S_i,\theta^0_{d(i)},\eta_1)\psi(S_j,\theta^0_{d(j)},\eta_2)-\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0)|\big] \\ \leq \; & \sup_{\eta_1\in\mathcal{T}}\|\varphi(S_i,\eta_1)-\varphi(S_i,\eta^0)\|_{P,2}\|\psi(S_j,\theta^0_{d(j)},\eta^0)\|_{P,2} \\ & \quad +\sup_{\eta_2\in\mathcal{T}}\|\psi(S_i,\theta^0_{d(i)},\eta^0)\|_{P,2}\|\varphi(S_j,\eta_2)-\varphi(S_j,\eta^0)\|_{P,2} \\ & \quad +\sup_{\eta_1,\eta_2\in\mathcal{T}}\|\varphi(S_i,\eta_1)-\varphi(S_i,\eta^0)\|_{P,2}\|\varphi(S_j,\eta_2)-\varphi(S_j,\eta^0)\|_{P,2} \end{aligned}$$

with P-probability at least $1-\Delta_N$ due to Hölder's inequality. Because all terms above are uniformly bounded due to Lemma 2.F.1, we infer that the

second set of summands in (2.22) is of order $o_P(1)$ due to Chernozhukov et al. (2018, Lemma 6.1). Finally, we bound the third set of summands in (2.22). Let $\varepsilon > 0$. We have

$$\begin{aligned} & P\Big(\Big|\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}\big(\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0) \\ & \quad - \mathbb{E}[\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0)]\big)\Big|^2 > \varepsilon^2\Big) \\ \leq\ & \tfrac{1}{\varepsilon^2 N^2}\Big(\textstyle\sum_{\{i,j\}\in E_D}\operatorname{Var}\big(\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_i,\theta^0_{d(j)},\eta^0)\big) \\ & \quad + \textstyle\sum_{\{i,j\},\{m,r\}\in E_D,\text{unequal}}\operatorname{Cov}\big(\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_i,\theta^0_{d(j)},\eta^0), \\ & \quad \psi(S_i,\theta^0_{d(m)},\eta^0)\psi(S_i,\theta^0_{d(r)},\eta^0)\big)\Big). \end{aligned} \tag{2.23}$$

Due to Lemma 2.F.1, the variance and covariance terms in (2.23) are uniformly bounded by constants. Furthermore, the covariance terms do only not vanish if S_i depends on S_m or S_r, or if S_j depends on S_m or S_r. In order to better describe these dependency relationships, we build a graph on the edge set of the dependency graph. We consider the graph $G' = (V', E')$ with $V' = E_D$ and such that an edge $\{\{i,j\},\{m,r\}\} \in E'$ if and only if at lease one of $\{i,m\}$, $\{i,r\}$, $\{j,m\}$, $\{j,r\}$ belongs to E_D. Consequently, $\{\{i,j\},\{m,r\}\} \in E'$ if and only if $(S_i, S_j) \not\perp\!\!\!\perp (S_m, S_r)$, in which case the covariance term in (2.23) corresponding to $\{i,j\}$ and $\{m,r\}$ does not vanish. Furthermore, we have $|E'| = 1/2|E_D|d'_{\max}$, where $d'_{\max}$ denotes the maximal degree of a node in G'. We have $d'_{\max} \leq 2d_{\max}$. Consequently, we have

$$\begin{aligned} & P\Big(\Big|\tfrac{1}{N}\textstyle\sum_{\{i,j\}\in E_D}\big(\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0) \\ & \quad - \mathbb{E}[\psi(S_i,\theta^0_{d(i)},\eta^0)\psi(S_j,\theta^0_{d(j)},\eta^0)]\big)\Big|^2 > \varepsilon^2\Big) \\ \leq\ & \tfrac{1}{\varepsilon^2 N^2}(|E_D| + |E'|)O(1) \\ =\ & \tfrac{1}{\varepsilon^2 N^2}(N d_{\max} + N d^2_{\max})O(1) \\ =\ & \tfrac{1}{\varepsilon^2 N}(o(N^{1/4}) + o(N^{1/2}))O(1) \\ =\ & o(1) \end{aligned}$$

due to Assumption 2.A.2. Therefore, we have established the statement of the present lemma because we have verified that all three sets of summands in (2.22) are of order $o_P(1)$. □

Proof of Theorem 2.2.6. The proof follows from Lemma 2.F.3 and 2.F.4. □

2.G | Extension to Estimate Global Effects

So far, we focused on the EATE, which is a direct effect. We intervened on each individual unit and left the treatment assignments of the other units as

they were.

Subsequently, we consider another type of treatment effect where we assess the effect of a single intervention that intervenes on all subjects simultaneously. Instead of the EATE in (2.4), we subsequently consider the global average treatment effect (GATE) with respect to the binary vector $\pi \in \{0,1\}^N$ of treatment assignments

$$\xi_N^0(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i^{do(\boldsymbol{W}=\pi)} - Y_i^{do(\boldsymbol{W}=1-\pi)}\right], \tag{2.24}$$

where $\boldsymbol{W} = (W_1, \ldots, W_N)$ denotes the complete vector of treatment assignments of all units. In practice, the most common choice is where all components of π equal 1. That is, the treatment effect comes from comparing the situation where all units are assigned to the treatment versus where no-one gets the treatment.

We use the same definition for S_i, $i \in [N]$ as before and denote the dependency graph on S_i, $i \in [N]$ by $G_D = (V, E_D)$. Furthermore, we let $\alpha(i) = \{j \in [N] \colon \{i,j\} \in E_D\} \cup \{i\}$ for $i \in [N]$ denote the nodes that share an edge with i in the dependency graph together with i itself. For some real number $\xi \in \mathbb{R}$ and a nuisance function triple $\eta = (g_1, g_0, h)$, consider the score function

$$\begin{aligned}\psi(S_i, \theta, \xi) \;=\; & g_1(C_i, X_i) - g_0(C_i, X_i) + \left(\Pi_{j\in\alpha(i)} \tfrac{W_j}{h(C_j,Z_j)}\right)\big(Y_i - g_1(C_i, X_i)\big) \\ & - \left(\Pi_{j\in\alpha(i)} \tfrac{1-W_j}{1-h(C_j,Z_j)}\right)\big(Y_i - g_0(C_i, X_i)\big) - \xi.\end{aligned} \tag{2.25}$$

In contrast to the score that we used for the EATE, this score includes additional factors $\frac{W_j}{h(C_j,Z_j)}$ and $\frac{1-W_j}{1-h(C_j,Z_j)}$ for units j that share an edge with i in the dependency graph. With the GATE, when we globally intervene on all treatment assignments at the same time, this also influences the X_i that are present in g_1 and g_0. In the score (2.25), the "correction terms" $(\Pi_{j\in\alpha(i)} \frac{W_j}{h(C_j,Z_j)})(Y_i - g_1(C_i, X_i))$ and $(\Pi_{j\in\alpha(i)} \frac{1-W_j}{1-h(C_j,Z_j)})(Y_i - g_0(C_i, X_i))$ are only active if i and the units from which it receives spillover effects have the same observed treatment assignment.

Let us denote by

$$\xi_i^0 = \mathbb{E}\left[Y_i^{do(\boldsymbol{W}=\pi)} - Y_i^{do(\boldsymbol{W}=1-\pi)}\right] = \mathbb{E}\left[g_1^0(C_i, X_i^\pi) - g_0^0(C_i, X_i^{1-\pi})\right]$$

the ith contribution in (2.24). Here,

$$X_i^\pi = \Big(f_x^1\big(\{(\pi_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big), \ldots, f_x^r\big(\{(\pi_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big)\Big)$$

denotes the feature vector where W_j is replaced by π_j, and

$$X_i^{1-\pi} = \Big(f_x^1\big(\{(1-\pi_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big), \ldots, f_x^r\big(\{(1-\pi_j, C_j)\}_{j\in[N]\setminus\{i\}}, G\big)\Big)$$

denotes the feature vector where W_j is replaced by $1-\pi_j$. The features Z_i^π and $Z_i^{1-\pi}$ are defined analogously. Similarly to Lemma 2.2.2, it can be shown that $\mathbb{E}[\psi(S_i, \xi_i^0, \eta^0)] = 0$ holds, which lets us identify the global treatment effect ξ_N^0 by

$$\xi_N^0 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\varphi(S_i, \eta^0)],$$

where

$$\begin{aligned}\varphi(S_i, \eta) \;=\; & g_1(C_i, X_i) - g_0(C_i, X_i) + \Big(\textstyle\prod_{j\in\alpha(i)} \frac{W_j}{h(C_j, Z_j)}\Big)\big(Y_i - g_1(C_i, X_i)\big) \\ & - \Big(\textstyle\prod_{j\in\alpha(i)} \frac{1-W_j}{1-h(C_j, Z_j)}\Big)\big(Y_i - g_0(C_i, X_i)\big).\end{aligned}$$

To estimate ξ_N^0, we apply the same procedure as for the ATE. The only difference is that when we evaluate the machine learning estimates, we do not use the observed treatment assignments, but instead insert the respective components of π and $1-\pi$. However, we insert the actually observed treatment assignments in the product terms $\prod_{j\in\alpha(i)} \frac{W_j}{h(C_j,Z_j)}$ and $\prod_{j\in\alpha(i)} \frac{1-W_j}{1-h(C_j,Z_j)}$. This gives the estimator $\hat{\xi}$. Analogously to Theorem 2.2.5 for the EATE, also the GATE with respect to π converges at the parametric rate and follows a Gaussian distribution asymptotically.

Theorem 2.G.1 (Asymptotic distribution of $\hat{\xi}$)**.** *Assume Assumption 3.B.2 (with θ replaced by ξ), 2.A.2, and 3.B.4 in the appendix in Section 3.B hold. Furthermore, assume that there exists a finite real constant L such that $|\alpha(i)| \leq L$ holds for all $i \in [N]$.*

Then, the estimator $\hat{\xi}$ of the GATE with respect to $\pi \in \{0,1\}^N$, ξ_N^0, satisfies

$$\sqrt{N}(\hat{\xi} - \xi_N^0) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty),$$

where σ_∞ is characterized in Assumption 2.A.3 with the ψ in (2.25)*. The convergence in* (2.G.1) *is in fact uniformly over the law P of the observations.*

This theorem requires that the number of spillover effects a unit receives is bounded. Theorem 2.2.5 that establishes the parametric convergence rate and asymptotic Gaussian distribution of the EATE estimator did not require such an assumption. The reason is that $h^0(C_i, Z_i)$ represents the conditional expectation of W_i given C_i and Z_i and consequently a probability taking

values in the interval $(0, 1)$. If we allowed $|\alpha(i)|$ to grow with N, the products $\Pi_{j\in\alpha(i)} \frac{W_j}{h(C_j,Z_j)}$ and $\Pi_{j\in\alpha(i)} \frac{1-W_j}{1-h(C_j,Z_j)}$ would diverge.

To estimate σ^2_∞ in Theorem 2.G.1, we can apply the procedure described in Section 2.2.5, where we replace ψ, φ, and the point estimators by the respective new quantities. Also an analogon of Theorem 2.2.6 holds, but where we assume the setting of Theorem 2.G.1 holds and that $|\mathcal{A}_d| \to \infty$ as $N \to \infty$ for all $d \geq 0$. In particular, we do not require Assumption 2.A.5 and 2.A.6 formulated in the appendix in Section 3.B. Furthermore, to prove consistency of the variance estimator, it is sufficient to establish that the degree-specific causal effect estimators $\hat{\xi}_d$, which are defined analogously to $\hat{\theta}_d$, are consistent. In particular, they are not required to converge at a particular rate.

Also van der Laan (2014), Sofrygin and van der Laan (2017), and Ogburn et al. (2022) consider semiparametric estimation of the GATE using TMLE. They also require a uniform bound of the number of spillover effects a unit receives to achieve the parametric convergence rate of their estimator. However, their methods cannot take into account spillover effects from more distant neighbors in the network than direct ones.

3 | Plugin Machine Learning for Partially Linear Mixed-Effects Models with Repeated Measurements

Joint work with

Peter Bühlmann

This chapter is based on the manuscript

C. Emmenegger and P. Bühlmann. Plugin machine learning for partially linear mixed-effects models with repeated measurements, 2021a. Preprint arXiv:2108.13657

Abstract

Traditionally, spline or kernel approaches in combination with parametric estimation are used to infer the linear coefficient (fixed effects) in a partially linear mixed-effects model for repeated measurements. Using machine learning algorithms allows us to incorporate complex interaction structures, nonsmooth terms, and high-dimensional variables. The linear variables and the response are adjusted nonparametrically for the nonlinear variables, and these adjusted variables satisfy a linear mixed-effects model in which the linear coefficient can be estimated with standard linear mixed-effects methods. We prove that the estimated fixed effects coefficient converges at the parametric rate, is asymptotically Gaussian distributed, and semiparametrically efficient. Two simulation studies demonstrate that our method outperforms a penalized regression spline approach in terms of coverage. We also illustrate our proposed approach on a longitudinal dataset with HIV-infected individuals. Software code for our method is available in the R-package `dmlalg`.

3.1 | Introduction

Repeated measurements data consists of observations from several experimental units, subjects, or groups under different conditions. This grouping or clustering of the individual responses into experimental units typically introduces dependencies: the different units are assumed to be independent, but there may be heterogeneity across units and correlation within units.

Mixed-effects models provide a powerful and flexible tool to analyze grouped data by incorporating fixed and random effects. Fixed effects are associated with the entire population, and random effects are associated with individual groups and model the heterogeneity across them and the dependence structure within them (Pinheiro and Bates, 2000). Linear mixed-effects models (Laird and Ware, 1982; Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000; Demidenko, 2004) impose a linear relationship between all covariates and the response. Partially linear mixed-effects models (Zeger and Diggle, 1994) extend the linear ones.

We consider the partially linear mixed-effects model

$$\mathbf{Y}_i = \mathbf{X}_i\beta_0 + g(\mathbf{W}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{3.1}$$

for groups $i \in \{1, \ldots, N\}$. There are n_i observations per group i. The unobserved random variable $\mathbf{b}_i$, called random effect, introduces correlation within its group i because all n_i observations within this group share $\mathbf{b}_i$. We make the assumption generally made that both the random effect $\mathbf{b}_i$ and the error term $\boldsymbol{\varepsilon}_i$ follow a Gaussian distribution (Pinheiro and Bates, 2000). The matrices $\mathbf{Z}_i$ assigning the random effects to group-level observations are fixed. The linear covariables $\mathbf{X}_i$ and the nonparametric and potentially high-dimensional covariables $\mathbf{W}_i$ are observed and random, and they may have dependent columns. Furthermore, the nonparametric covariables may contain nonlinear transformations and interaction terms of the linear ones. Please see Assumption 3.2.1 in Section 3.2 for further details.

Our aim is to estimate and make inference for the so-called fixed effect β_0 in (3.1) in the presence of a highly complex g using general machine learning algorithms. The parametric component β_0 provides a simple summary of the covariate effects that are of main scientific interest. The nonparametric component g enhances model flexibility because time trends and further covariates with possibly nonlinear and interaction effects can be modeled nonparametrically.

Repeated measurements, or longitudinal, data is omnipresent in empirical research. For example, assume we want to study the effect of a treatment over time. Observing the same subjects repeatedly presents three main advantages over having cross-sectional data. First, subjects can serve as their own controls. Second, the between-subject variability is explicitly modeled and can be excluded from the experimental error. This yields more efficient estimators of the relevant model parameters. Third, data can be collected more reliably (Davis, 2002; Fitzmaurice et al., 2011).

Various approaches have been considered in the literature to estimate the nonparametric component g in (3.1): kernel methods (Hart and Wehrly, 1986; Zeger and Diggle, 1994; Taavoni and Arashi, 2021b; Chen and Cao, 2017),

backfitting (Zeger and Diggle, 1994; Taavoni and Arashi, 2021b), spline methods (Rice and Silverman, 1991; Zhang, 2004; Guoyou and Zhongyi, 2007, 2009; Li and Zhu, 2010; Kim et al., 2017; Aniley et al., 2019), and local linear regression (Taavoni and Arashi, 2021b; Liang, 2009).

Our aim is to make inference for β_0 in the presence of potentially highly complex effects of $\mathbf{W}_i$ on $\mathbf{X}_i$ and $\mathbf{Y}_i$. First, we adjust $\mathbf{X}_i$ and $\mathbf{Y}_i$ for $\mathbf{W}_i$ by regressing $\mathbf{W}_i$ out of them using machine learning algorithms. These machine learning algorithms may yield biased results, especially if regularization methods are used, like for instance with the lasso (Tibshirani, 1996). Second, we fit a linear mixed-effects model to these regression residuals to estimate β_0. Our estimator of β_0 converges at the optimal $1/\sqrt{N}$ rate, follows a Gaussian distribution asymptotically, and is semiparametrically efficient.

We adapt double machine learning techniques of Chernozhukov et al. (2018) to estimate β_0 using general machine learning algorithms. To the best of our knowledge, this is the first work to allow the nonparametric nuisance components of a partially linear mixed-effects model to be estimated with arbitrary machine learners like random forests (Breiman, 2001) or the lasso (Tibshirani, 1996; Bühlmann and van de Geer, 2011). In contrast to the setting and proofs of Chernozhukov et al. (2018), we have dependent data and need to incorporate this accordingly. Chernozhukov et al. (2018) introduce double machine learning and develop estimation of the low-dimensional linear regression parameter vector in a partially linear model. Their estimator converges at the parametric rate and is asymptotically Gaussian due to Neyman orthogonality and sample splitting with cross-fitting. We would like to remark that nonparametric nuisance 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 (1997). However, these conditions limit the complexity of the function class, and machine learning algorithms usually do not satisfy them. Particularly, these conditions fail to hold if the dimension of the nonparametric variables increases with the sample size (Chernozhukov et al., 2018). We show that the desirable properties of double machine learning also hold in the context of partially linear mixed-effects models: such a further development of plug-in machine learning methods is nontrivial and practically highly relevant.

3.1.1 | Additional Literature

Expositions and overviews of mixed-effects modeling techniques can be found in Pinheiro (1994); Davidian and Giltinan (1995); Vonesh and Chinchilli (1997); Pinheiro and Bates (2000); Davidian and Giltinan (2003).

Zhang et al. (1998) consider partially linear mixed-effects models and estimate the nonparametric component with natural cubic splines. They treat the

smoothing parameter as an extra variance component that is jointly estimated with the other variance components of the model. Masci et al. (2019) consider partially linear mixed-effects models for unsupervised classification with discrete random effects. Schelldorfer et al. (2011) consider high-dimensional linear mixed-effects models where the number of fixed effects coefficients may be much larger than the overall sample size. To estimate and make inference for the first, say, d components of the linear coefficient in such a high-dimensional mixed-effects model, our approach may consider the remaining components as an additive contribution $\mathbf{W}_i\beta_{0,-(1:d)}$ in the model and may adjust for them using the lasso (Tibshirani, 1996). Debiased fixed effects estimators in high-dimensional linear mixed effects models are studied by Li et al. (2021) and Bradic et al. (2020). Taavoni and Arashi (2021a) employ a regularization approach in generalized partially linear mixed-effects models using regression splines to approximate the nonparametric component. Wood and Scheipl (2020) use penalized regression splines where the penalized components are treated as random effects.

The unobserved random variables in the partially linear mixed-effects model in (3.1) are assumed to follow a Gaussian distribution. Taavoni et al. (2021) introduce multivariate t partially linear mixed-effects models for longitudinal data. They consider t-distributed random effects to account for outliers in the data. Fahrmeir and Kneib (2011, Chapter 4) relax the assumption of Gaussian random effects in generalized linear mixed models. They consider nonparametric Dirichlet processes and Dirichlet process mixture priors for the random effects. Ohinata (2012, Chapter 3) consider partially linear mixed-effects models and make no distributional assumptions for the random terms, and the nonparametric component is estimated with kernel methods. Lu (2016) consider a partially linear mixed-effects model that is nonparametric in time and that features asymmetrically distributed errors and missing data.

Furthermore, methods have been developed to analyze repeated measurements data that are robust to outliers. Guoyou and Zhongyi (2008) consider robust estimating equations and estimate the nonparametric component with a regression spline. Tang et al. (2015) consider median-based regression methods in a partially linear model with longitudinal data to account for highly skewed responses. Lin et al. (2018) present an estimation technique in partially linear models for longitudinal data that is doubly robust in the sense that it simultaneously accounts for missing responses and mismeasured covariates.

It is prespecified in the partially linear mixed-effects model (3.1) which covariates are modeled with random effects. Simultaneous variable selection for fixed effects variables and random effects has been developed by Bondell et al. (2010); Ibrahim et al. (2011). They use penalized likelihood approaches. Li and Zhu (2010) use a nonparametric test to test the existence of random effects in

partially linear mixed-effects models. Zhang and Xue (2020) propose a variable selection procedure for the linear covariates of a generalized partially linear model with longitudinal data.

Outline of the Paper. Section 3.2 presents our plug-in machine learning estimator of the linear coefficient in a partially linear mixed-effects model. Section 3.3 presents our numerical results. Proofs and technical assumptions are presented in the appendix.

Notation. We denote by $[N]$ the set $\{1, 2, \ldots, N\}$. We add the probability law P 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 for $p \geq 1$ 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 $\xrightarrow{\mathcal{L}}$ convergence in distribution. The symbol $\perp\!\!\!\perp$ denotes independence of random variables. We denote by $\mathbb{1}_n$ the $n \times n$ identity matrix and omit the subscript n if we do not want to emphasize the dimension. We denote the d-variate Gaussian distribution by $\mathcal{N}_d$.

3.2 | Model Formulation and the Plug-in Machine Learning Estimator

We consider repeated measurements data that is grouped according to experimental units or subjects. This grouping structure introduces dependency in the data. The individual experimental units or groups are assumed to be independent, but there may be some between-group heterogeneity and within-group correlation. We consider the partially linear mixed-effects model

$$\mathbf{Y}_i = \mathbf{X}_i\beta_0 + g(\mathbf{W}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i \in [N] \tag{3.2}$$

for groups i as in (3.1) to model the between-group heterogeneity and within-group correlation with random effects. We have n_i observations per group that are concatenated row-wise into $\mathbf{Y}_i \in \mathbb{R}^{n_i}$, $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$, and $\mathbf{W}_i \in \mathbb{R}^{n_i \times v}$. The nonparametric variables may be high-dimensional, but d is fixed. Both $\mathbf{X}_i$ and $\mathbf{W}_i$ are random. The $\mathbf{X}_i$ and $\mathbf{W}_i$ belonging to the same group i may be dependent. For groups $i \neq j$, we assume $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$, $\mathbf{W}_i \perp\!\!\!\perp \mathbf{W}_j$, and $\mathbf{X}_i \perp\!\!\!\perp \mathbf{W}_j$. Moreover, we assume that all within-unit observations of the linear and nonlinear covariates, namely $((\mathbf{X}_i)_{t,\cdot}, (\mathbf{W}_i)_{t,\cdot})$ for all $i \in [N]$ and all $t \in [n_i]$, are independent and identically distributed. We assume that $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$ is fixed. The random variable $\mathbf{b}_i \in \mathbb{R}^q$ denotes a group-specific vector of random regression coefficients that is assumed to follow a Gaussian distribution. The dimension q of the random effects model is fixed. Also the

error terms are assumed to follow a Gaussian distribution as is commonly done in a mixed-effects models framework (Pinheiro and Bates, 2000). All groups i share the common linear coefficient β_0 and the potentially complex function $g\colon \mathbb{R}^v \to \mathbb{R}$. The function g is applied row-wise to $\mathbf{W}_i$, denoted by $g(\mathbf{W}_i)$.

We denote the total number of observations by $N_T := \sum_{i=1}^{N} n_i$. We assume that the numbers n_i of within-group observations are uniformly upper bounded by $n_{\max} < \infty$. Asymptotically, the number of groups, N, goes to infinity.

Our distributional and independency assumptions are summarized as follows:

Assumptions 3.2.1. *Consider the partially linear mixed-effects model* (3.2). *We assume that there is some $\sigma_0 > 0$ and some symmetric positive definite matrix $\Gamma_0 \in \mathbb{R}^{q\times q}$ such that the following conditions hold.*

3.2.1.1 The random effects $\mathbf{b}_1, \ldots, \mathbf{b}_N$ are independent and identically distributed $\mathcal{N}_q(\mathbf{0}, \Gamma_0)$.

3.2.1.2 The error terms $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are independent and follow a Gaussian distribution, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_0^2 \mathbb{1}_{n_i})$ for $i \in [N]$, with the common variance component σ_0^2.

3.2.1.3 The variables $\mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are independent.

3.2.1.4 For all $i, j \in [N]$, $i \neq j$, we have $(\mathbf{b}_i, \boldsymbol{\varepsilon}_i) \perp\!\!\!\perp (\mathbf{W}_i, \mathbf{X}_i)$ and $(\mathbf{b}_i, \boldsymbol{\varepsilon}_i) \perp\!\!\!\perp (\mathbf{W}_j, \mathbf{X}_j)$.

3.2.1.5 For all $i \in [N]$ and all $t \in [n_i]$, we have that $((\mathbf{X}_i)_{t,\cdot}, (\mathbf{W}_i)_{t,\cdot})$ are independent and identically distributed.

We would like to remark that the distribution of the error terms $\boldsymbol{\varepsilon}_i$ in Assumption 3.2.1.2 can be generalized to $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_0^2 \Lambda_i(\boldsymbol{\lambda}))$, where $\Lambda_i(\boldsymbol{\lambda}) \in \mathbb{R}^{n_i \times n_i}$ is a symmetric positive definite matrix parametrized by some finite-dimensional parameter vector $\boldsymbol{\lambda}$ that all groups have in common. For the sake of notational simplicity, we restrict ourselves to Assumption 3.2.1.2.

Moreover, we may consider stochastic random effects matrices $\mathbf{Z}_i$. Alternatively, the nonparametric variables $\mathbf{W}_i$ may be part of the random effects matrix. In this case, we consider the random effects matrix $\overline{\mathbf{Z}}_i = \zeta(\mathbf{Z}_i, \mathbf{W}_i)$ for some known function ζ in (3.2) instead of $\mathbf{Z}_i$. Please see Section 3.D in the appendix for further details. For simplicity, we restrict ourselves to fixed random effects matrices $\mathbf{Z}_i$ that are disjoint from $\mathbf{W}_i$.

The unknown parameters in our model are β_0, Γ_0, and σ_0. Our aim is to estimate β_0 and make inference for it. Although the variance parameters Γ_0 and σ_0 need to be estimated consistently to construct an estimator of β_0, it is not our goal to perform inference for them.

3.2.1 | The Plug-in Machine Learning Estimator

Subsequently, we describe our plug-in machine learning estimator of β_0 in (3.2). To motivate our procedure, we first consider the population version with the residual terms

$$\mathbf{R}_{\mathbf{X}_i} := \mathbf{X}_i - \mathbb{E}[\mathbf{X}_i|\mathbf{W}_i] \quad \text{and} \quad \mathbf{R}_{\mathbf{Y}_i} := \mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i|\mathbf{W}_i] \quad \text{for} \quad i \in [N]$$

that adjust $\mathbf{X}_i$ and $\mathbf{Y}_i$ for $\mathbf{W}_i$. On this adjusted level, we have the linear mixed-effects model

$$\mathbf{R}_{\mathbf{Y}_i} = \mathbf{R}_{\mathbf{X}_i}\beta_0 + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i \in [N] \tag{3.3}$$

due to (3.2) and Assumption 3.2.1.4. In particular, the adjusted and grouped responses in this model are independent in the sense that we have $\mathbf{R}_{\mathbf{Y}_i} \perp\!\!\!\perp \mathbf{R}_{\mathbf{Y}_j}$ for $i \neq j$. The strategy now is to first estimate the residuals with machine learning algorithms and then use linear mixed model techniques to infer β_0. This is done with sample splitting and cross-fitting, and the details are described next.

Let us define $\Sigma_0 := \sigma_0^{-2}\Gamma_0$ and $\mathbf{V}_{0,i} := (\mathbf{Z}_i\Sigma_0\mathbf{Z}_i^T + \mathbb{1}_{n_i})$ so that we have

$$(\mathbf{R}_{\mathbf{Y}_i}|\mathbf{W}_i, \mathbf{X}_i) \sim \mathcal{N}_{n_i}\big(\mathbf{R}_{\mathbf{X}_i}\beta_0, \sigma_0^2\mathbf{V}_{0,i}\big). \tag{3.4}$$

We assume that there exist functions $m_X^0 \colon \mathbb{R}^v \to \mathbb{R}^d$ and $m_Y^0 \colon \mathbb{R}^v \to \mathbb{R}$ that we can apply row-wise to $\mathbf{W}_i$ to have $\mathbb{E}[\mathbf{X}_i|\mathbf{W}_i] = m_X^0(\mathbf{W}_i)$ and $\mathbb{E}[\mathbf{Y}_i|\mathbf{W}_i] = m_Y^0(\mathbf{W}_i)$, which is conceivable due to Assumption 3.2.1.5. In particular, m_X^0 and m_Y^0 do not depend on the grouping index i. Let $\eta^0 := (m_X^0, m_Y^0)$ denote the true unknown nuisance parameter. Let us denote by $\theta_0 := (\beta_0, \sigma_0^2, \Sigma_0)$ the complete true unknown parameter vector and by $\theta := (\beta, \sigma^2, \Sigma)$ and $\mathbf{V}_i := \mathbf{Z}_i\Sigma\mathbf{Z}_i^T + \mathbb{1}_{n_i}$ respective general parameters. The log-likelihood of group i is given by

$$\begin{aligned}\ell_i\big(\theta, \eta^0\big) \;=\; & -\tfrac{n_i}{2}\log(2\pi) - \tfrac{n_i}{2}\log(\sigma^2) - \tfrac{1}{2}\log\big(\det(\mathbf{V}_i)\big) \\ & -\tfrac{1}{2\sigma^2}(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i}\beta)^T\mathbf{V}_i^{-1}(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i}\beta) - \log\big(p(\mathbf{W}_i, \mathbf{X}_i)\big),\end{aligned} \tag{3.5}$$

where $p(\mathbf{W}_i, \mathbf{X}_i)$ denotes the joint density of $\mathbf{W}_i$ and $\mathbf{X}_i$. We assume that $p(\mathbf{W}_i, \mathbf{X}_i)$ does not depend on θ. The true nuisance parameter η^0 in the log-likelihood (3.5) is unknown and estimated with machine learning algorithms (see below). Denote by $\eta := (m_X, m_Y)$ some general nuisance parameter. The terms that adjust $\mathbf{X}_i$ and $\mathbf{Y}_i$ for $\mathbf{W}_i$ with this general nuisance parameter are given by $\mathbf{X}_i - m_X(\mathbf{W}_i)$ and $\mathbf{Y}_i - m_Y(\mathbf{W}_i)$. Up to additive constants that do not depend on θ and η, we thus consider maximum likelihood estimation with

the likelihood

$$\ell_i(\theta,\eta) = -\tfrac{n_i}{2}\log(\sigma^2) - \tfrac{1}{2}\log\big(\det(\mathbf{V}_i)\big) \\ -\tfrac{1}{2\sigma^2}\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big)^T \mathbf{V}_i^{-1} \\ \cdot\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big),$$

which is a function of both the finite-dimensional parameter θ and the infinite-dimensional nuisance parameter η.

Our estimator of β_0 is constructed as follows adapting double machine learning techniques. We estimate η^0 with machine learning algorithms and plug these estimators into the estimating equations for θ_0, equation (3.6) below, to obtain an estimator for β_0. This procedure is done with sample splitting and cross-fitting as explained next.

Consider repeated measurements from N experimental units, subjects, or groups as in (3.2). Denote by $\mathbf{S}_i := (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)$ the observations of group i. First, we split the group indices $[N]$ into $K \geq 2$ disjoint sets $I_1, \ldots, I_K$ of approximately equal size in the sense that the number of unit-level observations belonging to each set are asymptotically of the same order. The number of observations per unit may differ, but is assumed to be uniformly bounded. That is, we avoid too unbalanced settings. Please see Section 3.B in the appendix for further details.

For each $k \in [K]$, we estimate the conditional expectations $m_X^0(W)$ and $m_Y^0(W)$ with data from I_k^c. We call the resulting estimators $\hat{m}_X^{I_k^c}$ and $\hat{m}_Y^{I_k^c}$, respectively. Then, the adjustments $\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k} := \mathbf{X}_i - \hat{m}_X^{I_k^c}(\mathbf{W}_i)$, and $\widehat{\mathbf{R}}_{\mathbf{Y}_i}^{I_k} := \mathbf{Y}_i - \hat{m}_Y^{I_k^c}(\mathbf{W}_i)$ for $i \in I_k$ are evaluated on I_k, the complement of I_k^c. Let $\hat{\eta}^{I_k^c} := (\hat{m}_X^{I_k^c}, \hat{m}_Y^{I_k^c})$ denote the estimated nuisance parameter. Consider the score function $\psi(\mathbf{S}_i; \theta, \eta) := \nabla_\theta \ell_i(\theta, \eta)$, where ∇_θ denotes the gradient with respect to θ interpreted as a vector. On each set I_k, we consider an estimator $\hat{\theta}_k = (\hat{\beta}_k, \hat{\sigma}_k^2, \hat{\Sigma}_k)$ of θ_0 that, approximately, in the sense of Assumption 3.B.3.3 in the appendix, solves

$$\frac{1}{n_{T,k}} \sum_{i \in I_k} \psi\big(\mathbf{S}_i; \hat{\theta}_k, \hat{\eta}^{I_k^c}\big) = \frac{1}{n_{T,k}} \sum_{i \in I_k} \nabla_\theta \ell_i(\theta, \eta) \stackrel{!}{=} \mathbf{0}, \tag{3.6}$$

where $n_{T,k} := \sum_{i \in I_k} n_i$ denotes the total number of observations from experimental units that belong to the set I_k. These K estimators $\hat{\theta}_k$ for $k \in [K]$ are assembled to form the final cross-fitting estimator

$$\hat{\beta} := \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k \tag{3.7}$$

of β_0. We remark that one can simply use linear mixed model computation and software to compute $\hat{\beta}_k$ based on the estimated residuals $\widehat{\mathbf{R}}^{I_k}$. The estimator $\hat{\beta}$ fundamentally depends on the particular sample split. To alleviate this effect, the overall procedure may be repeated $\mathcal{S}$ times (Chernozhukov et al., 2018). The $\mathcal{S}$ point estimators are aggregated by the median, and an additional term accounting for the random splits is added to the variance estimator of $\hat{\beta}$; please see Algorithm 2 that presents the complete procedure.

Algorithm 2: Plug-in machine learning for partially linear mixed-effects models with repeated measurements.

Input : N grouped observations $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)\}_{i\in[N]}$ from model (3.2) satisfying Assumption 3.2.1, a natural number K, a natural number $\mathcal{S}$.

Output : An estimator of β_0 in (3.2) together with its estimated asymptotic variance.

for $s \in [\mathcal{S}]$ **do**
- Split the grouped observation index set $[N]$ into K sets $I_1, \ldots, I_K$ of approximately equal size.
- **for** $k \in K$ **do**
  - Compute the conditional expectation estimators $\hat{m}_X^{I_k^c}$ and $\hat{m}_Y^{I_k^c}$ with some machine learning algorithm and data from I_k^c.
  - Evaluate the adjustments $\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k} = \mathbf{X}_i - \hat{m}_X^{I_k^c}(\mathbf{W}_i)$ and $\widehat{\mathbf{R}}_{\mathbf{Y}_i}^{I_k} = \mathbf{Y}_i - \hat{m}_Y^{I_k^c}(\mathbf{W}_i)$ for $i \in I_k$.
  - Compute $\hat{\theta}_{k,s} = (\hat{\beta}_{k,s}, \hat{\sigma}^2_{k,s}, \hat{\Sigma}_{k,s})$ using, for instance, linear mixed model techniques.
- **end**
- Compute $\hat{\beta}_s = \frac{1}{K}\sum_{k=1}^K \hat{\beta}_{k,s}$ as an approximate solution to (3.6).
- Compute an estimate $\hat{T}_{0,s}$ of the asymptotic variance-covariance matrix T_0 in Theorem 3.2.2.

end

Compute $\hat{\beta} = \text{median}_{s\in[\mathcal{S}]}(\hat{\beta}_s)$.

Estimate T_0 by $\hat{T}_0 = \text{median}_{s\in[\mathcal{S}]}(\hat{T}_{0,s} + (\hat{\beta} - \hat{\beta}_s)(\hat{\beta} - \hat{\beta}_s)^T)$.

3.2.2 | Theoretical Properties of the Plug-in Machine Learning Estimator

The estimator $\hat{\beta}$ as in (3.7) converges at the parametric rate, is asymptotically Gaussian distributed, and semiparametrically efficient.

Theorem 3.2.2. *Consider grouped observations* $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Y}_i)\}_{i\in[N]}$ *from the partially linear mixed-effects model* (3.2) *that satisfy Assumption 3.2.1 such that* $p(\mathbf{W}_i, \mathbf{X}_i)$ *does not depend on* θ*. Let* $N_T := \sum_{i=1}^{N} n_i$ *denote the total number of unit-level observations. Furthermore, suppose the assumptions in Section 3.B in the appendix hold, and consider the symmetric positive-definite matrix* T_0 *given in Assumption 3.B.2.8 in the appendix. Then,* $\hat{\beta}$ *as in* (3.7) *concentrates in a* $1/\sqrt{N_T}$ *neighborhood of* β_0*, is centered Gaussian, namely*

$$\sqrt{N_T} T_0^{\frac{1}{2}} (\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_d(\mathbf{0}, \mathbb{1}_d) \quad (N \to \infty), \tag{3.8}$$

and semiparametrically efficient. The convergence in (3.8) *is in fact uniformly over the law* P *of* $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Y}_i)\}_{i\in[N]}$.

Please see Section 3.C.4 in the appendix for a proof of Theorem 3.2.2. Our proof builds on Chernozhukov et al. (2018), but we have to take into account the correlation within units that is introduced by the random effects.

The inverse asymptotic variance-covariance matrix T_0 can be consistently estimated; see Lemma 3.C.18 in the appendix. Semiparametric efficiency follows from Lin and Carroll (2001, Section 5).

The assumptions in Section 3.B of the appendix specify regularity conditions and required convergence rates of the machine learning estimators. The machine learning errors need to satisfy the product relationship

$$\|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2} \big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2} \big) \ll N^{-\frac{1}{2}}.$$

This bound requires that only the products of the machine learning estimation errors $\|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}$ and $\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2}$ but not the individual ones need to vanish at a rate smaller than $N^{-1/2}$. In particular, the individual estimation errors may vanish at the rate smaller than $N^{-1/4}$. This is achieved by many machine learning methods (cf. Chernozhukov et al. (2018)): ℓ_1-penalized and related methods in a variety of sparse models (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Belloni et al., 2011; Belloni and Chernozhukov, 2011; Belloni et al., 2012; Belloni and Chernozhukov, 2013), forward selection in sparse models (Kozbur, 2020), L_2-boosting in sparse linear

models (Luo and Spindler, 2016), a class of regression trees and random forests (Wager and Walther, 2016), and neural networks (Chen and White, 1999).

We note that so-called Neyman orthogonality makes score functions insensitive to inserting potentially biased machine learning estimators of the nuisance parameters. A score function is Neyman orthogonal if its Gateaux derivative vanishes at the true θ_0 and the true η^0. In particular, Neyman orthogonality is a first-order property. The product relationship of the machine learning estimating errors described above is used to bound second-order terms. We refer to Section 3.C.4 in the appendix for more details.

3.3 | Numerical Experiments

Subsequently, we apply our plug-in machine learning method to an empirical and a pseudorandom dataset and in a simulation study. Our implementation is available in the R-package `dmlalg` (Emmenegger, 2021).

3.3.1 | Empirical Analysis: CD4 Cell Count Data

First, we apply our method to longitudinal CD4 cell counts data collected from human immunodeficiency virus (HIV) seroconverters. This data has previously been analyzed by Zeger and Diggle (1994) and is available in the R-package `jmcm` (Pan and Pan, 2017) as `aids`. It contains 2376 observations of CD4 cell counts measured on 369 subjects. The data was collected during a period ranging from 3 years before to 6 years after seroconversion. The number of observations per subject ranges from 1 to 12, but for most subjects, 4 to 10 observations are available. Please see Zeger and Diggle (1994) for more details on this dataset.

Apart from time, five other covariates are measured: the age at seroconversion in years (age), the smoking status measured by the number of cigarette packs consumed per day (smoking), a binary variable indicating drug use (drugs), the number of sex partners (sex), and the depression status measured on the Center for Epidemiologic Studies Depression (CESD) scale (cesd), where higher CESD values indicate the presence of more depression symptoms.

We incorporate a random intercept per person. Furthermore, we consider a square-root transformation of the CD4 cell counts to reduce the skewness of this variable as proposed by Zeger and Diggle (1994). The CD4 counts are our response. The covariates that are of scientific interest are considered as X's, and the remaining covariates are considered as W's in the partially linear mixed-effects model (3.2). The effect of time is modeled nonparametrically, but there are several options to model the other covariates. Other models than partially linear mixed-effects model have also been considered in the literature to analyze this dataset. For instance, Fan and Zhang (2000) consider a functional linear model where the linear coefficients are a function of the time.

	age	smoking	drugs	sex	cesd
$W = (\text{time})$	0.004	0.752	0.704	0.001	−0.042
	(0.027)	(0.123)	(0.360)	(0.043)	(0.015)
$W =$	-	0.620	0.602	-	−0.047
$(\text{time}, \text{age}, \text{sex})$		(0.126)	(0.335)		(0.015)
Zeger and	0.037	0.27	0.37	0.10	−0.058
Diggle (1994)	(0.18)	(0.15)	(0.31)	(0.038)	(0.015)
Taavoni and	$1.5 \cdot 10^{-17}$	0.152	0.130	0.0184	−0.0141
Arashi (2021b)	$(3.5 \cdot 10^{-17})$	(0.208)	(0.071)	(0.0039)	(0.0061)
Wang et al.	0.010	0.549	0.584	0.080	−0.045
(2005)	(0.033)	(0.144)	(0.331)	(0.038)	(0.013)
Guoyou and	0.006	0.538	0.637	0.066	−0.042
Zhongyi (2008)	(0.038)	(0.136)	(0.350)	(0.040)	(0.015)

Table 3.3.1: Estimates of the linear coefficient and its standard deviation in parentheses with our method for nonparametrically adjusting for time (first row) and for time, age, and sex (second row). The remaining rows display the results from Zeger and Diggle (1994, Section 5), Taavoni and Arashi (2021b, Table 1, "Kernel"), Wang et al. (2005, Table 2, "Semiparametric efficient scenario I"), and Guoyou and Zhongyi (2008, Table 5, "Robust"), respectively.

We consider two partially linear mixed-effects models for this dataset. First, we incorporate all covariates except time linearly. Most approaches in the literature employing a partially linear mixed-effects model for this data that model time nonparametrically report that sex and cesd are significant and that either smoking or drugs is significant as well; see for instance Zeger and Diggle (1994); Taavoni and Arashi (2021b); Wang et al. (2005). Guoyou and Zhongyi (2008) develop a robust estimation method for longitudinal data and estimate nonlinear effects from time with regression splines. With the CD4 dataset, they find that smoking and cesd are significant.

We apply our method with $K = 2$ sample splits, $\mathcal{S} = 100$ repetitions of splitting the data, and learn the conditional expectations with random forests that consist of 500 trees whose minimal node size is 5. Like Guoyou and Zhongyi (2008), we conclude that smoking and cesd are significant; please see the first row of Table 4.5.1 for a more precise account of our findings. Apart from sex, our point estimators are larger or of about the same size in absolute value as what Guoyou and Zhongyi (2008) obtain. However, apart from age, the standard deviations are slightly larger with our method. This can be expected because random forests are more complex than the regression splines Guoyou and Zhongyi (2008) employ.

We consider a second estimation approach where we model the variables time,

age, and sex nonparametrically and allow them to interact. It is conceivable that these variables are not (causally) influenced by smoking, drugs, and cesd and that they are therefore exogenous. The variables smoking, drugs, and cesd are modeled linearly, and they are considered as treatment variables. Some direct causal effect interpretations are possible if one is willing to assume, for instance, that the nonparametric adjustment variables are causal parents of the linear variables or the response. However, we do not pursue this line of thought further. We estimate the conditional expectations given the three nonparametric variables time, age, and sex again with random forests that consist of 500 trees whose minimal node size is 5 and use $K = 2$ and $\mathcal{S} = 100$ in Algorithm 2. We again find that smoking and cesd are significant; please see the second row of Table 4.5.1. This cannot be expected a priori because this second model incorporates more complex adjustments, which can lead to less significant variables.

3.3.2 | Pseudorandom Simulation Study: CD4 Cell Count Data

Subsequently, we consider the CD4 cell count data from the previous subsection and perform a pseudorandom simulation study. The variables smoking, drugs, and cesd are modeled linearly and the variables time, age, and sex nonparametrically. We condition on these six variables in our simulation. That is, they are the same in all repetitions. The function g in (3.2) is chosen as a regression tree that we built beforehand. We let $\beta_0 = (0.62, 0.6, -0.05)^T$, where the first component corresponds to smoking, the second one to drugs, and the last one to cesd, consider a standard deviation of the random intercept per subject of 4.36, and a standard deviation of the error term of 4.35. These are the point estimates of the respective quantities obtained in the previous subsection.

Our fitting procedure uses random forests consisting of 500 trees whose minimal node size is 5 to estimate the conditional expectations, and we use $K = 2$ and $\mathcal{S} = 10$ in Algorithm 2. We perform 5000 simulation runs. We compare the performance of our method with that of the spline-based function `gamm4` from the package `gamm4` (Wood and Scheipl, 2020) for the statistical software R (R Core Team, 2019). This method represents the nonlinear part of the model by smooth additive functions and estimates them by penalized regression splines. The penalized components are treated as random effects and the unpenalized components as fixed effects.

The results are displayed in Figure 3.3.1. With our method, `mmdml`, the two-sided confidence intervals for β_0 are of about the same length but achieve a coverage that is closer to the nominal 95% confidence level than with `gamm4`. The `gamm4` method largely undercovers the packs component of β_0, which can be explained by the incorporated bias.

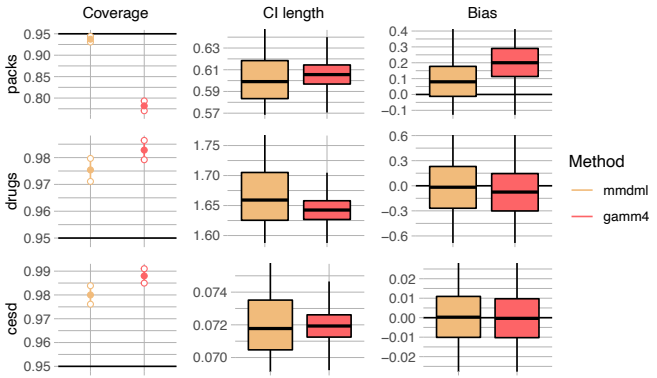


Figure 3.3.1: Coverage and length of two-sided confidence intervals at significance level 5% and bias for our method, `mmdml`, and `gamm4`. In the coverage plot, solid dots represent point estimators, and circles represent 95% confidence bands with respect to the 5000 simulation runs. The confidence interval length and bias are displayed with box plots without outliers.

3.3.3 | Simulation Study

Finally, we carry out a simulation study with a partially linear mixed-effects model with $q = 3$ random effects and where β_0 is 1-dimensional. Every subject has their own random intercept term and a nested random effect with two levels. Thus, the random effects structure is more complex than in the previous two subsections because these models only used a random intercept. We compare three data generating mechanisms: one where the function g is nonsmooth and the number of observations per group is balanced, one where the function g is smooth and the number of observations per group is balanced, and one where the function g is nonsmooth and the number of observations per group is unbalanced; please see Section 3.A in the appendix for more details.

We estimate the nonparametric nuisance components, that is, the conditional expectations, with random forests consisting of 500 trees whose minimal node size is 5. Furthermore, we use $K = 2$ and $\mathcal{S} = 10$ in Algorithm 2. We perform 1000 simulation runs and consider different numbers of groups N. As in the previous subsection, we compare the performance of our method with `gamm4`.

The results are displayed in Figure 3.3.2. Our method, `mmdml`, highly outperforms `gamm4` in terms of coverage for nonsmooth g because the coverage of `gamm4` equals 0 due to its substantial bias. Our method overcovers slightly due to the correction factor that results from the $\mathcal{S}$ repetitions. However, this correction factor is highly recommended in practice. With smooth g, `gamm4` is

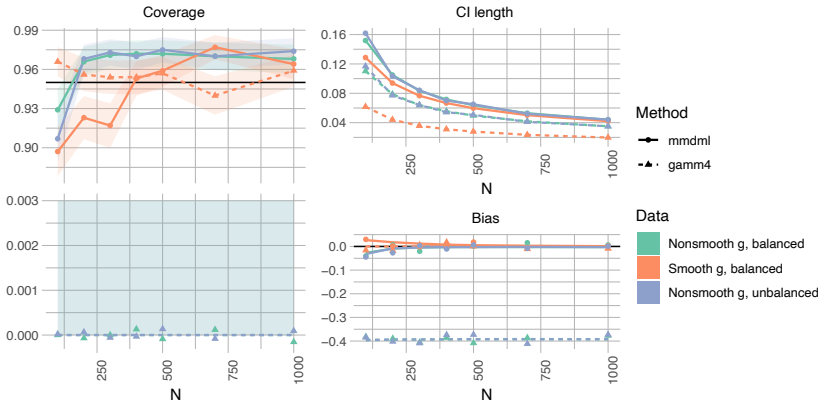


Figure 3.3.2: Coverage and median length of two-sided confidence intervals for β_0 at significance level 5% (true $\beta_0 = 0.5$) and median bias for three data generating scenarios for our method, `mmdml`, and `gamm4`. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1000 simulation runs. The dots in the coverage and bias plot are jittered, but neither are their interconnecting lines nor their confidence bands.

closer to the nominal coverage and has shorter confidence intervals than our method. Because the underlying model is smooth and additive, a spline-based estimator is better suited. In all scenarios, our method outputs longer confidence intervals than `gamm4` because we use random forests; consistent with theory, the difference in absolute value decreases though when N increases.

3.4 | Conclusion

Our aim was to develop inference for the linear coefficient β_0 of a partially linear mixed-effects model that includes a linear term and potentially complex nonparametric terms. Such models can be used to describe heterogeneous and correlated data that feature some grouping structure, which may result from taking repeated measurements. Traditionally, spline or kernel approaches are used to cope with the nonparametric part of such a model. We presented a plug-in machine learning scheme that adapts double machine learning techniques of Chernozhukov et al. (2018) to estimate any nonparametric components with arbitrary machine learning algorithms. This allowed us to consider complex nonparametric components with interaction structures and high-dimensional variables.

Our proposed method is as follows. First, the nonparametric variables are

regressed out from the response and the linear variables. This step adjusts the response and the linear variables for the nonparametric variables and may be performed with any machine learning algorithm. The adjusted variables satisfy a linear mixed-effects model, where the linear coefficient β_0 can be estimated with standard linear mixed-effects techniques. We showed that the estimator of β_0 asymptotically follows a Gaussian distribution, converges at the parametric rate, and is semiparametrically efficient. This asymptotic result allows us to perform inference for β_0.

Empirical experiments demonstrated the performance of our proposed method. We conducted an empirical and pseudorandom data analysis and a simulation study. The simulation study and the pseudorandom experiment confirmed the effectiveness of our method in terms of coverage, length of confidence intervals, and estimation bias compared to a penalized regression spline approach relying on additive models. In the empirical experiment, we analyzed longitudinal CD4 cell counts data collected from HIV-infected individuals. In the literature, most methods only incorporate the time component nonparametrically to analyze this dataset. Because we estimate nonparametric components with machine learning algorithms, we can allow several variables to enter the model nonlinearly, and we can allow these variables to interact.

The R-package `dmlalg` (Emmenegger, 2021) provides an implementation of our method.

Acknowledgements

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). We would like to thank Michael Law for constructive comments.

Appendix

3.A | Data Generating Mechanism for Simulation Study

Let $n = 15$. For all scenarios except the unbalanced one, we sample the number of observations for each experimental unit from $\{n-3, n-2, \ldots, n+2, n+3\}$ with equal probability. For the unbalanced scenario, we sample the number of observations for each experimental unit from $\{1, 2, \ldots, 2n-2, 2n-1\}$ with equal probability. We consider 3-dimensional nonparametric variables. For $w = (w_1, w_2, w_3) \in \mathbb{R}^3$, consider the real-valued functions

$$
\begin{aligned}
& h(w) \\
:= \; & -3 \cdot \mathbb{1}_{w_3>0}\mathbb{1}_{w_1>0} + 2 \cdot \mathbb{1}_{w_3>0}\mathbb{1}_{w_1\leq 0} - \mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3\leq -1} - 2 \cdot \mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3>-1}\mathbb{1}_{w_2>0} \\
& -3 \cdot \mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3>-1}\mathbb{1}_{w_2\leq 0}\mathbb{1}_{w_1>0.75} + \mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3>-1}\mathbb{1}_{w_2\leq 0}\mathbb{1}_{w_1\leq 0.75}
\end{aligned}
$$

and

$$
\begin{aligned}
& g(w) \\
:= \; & \mathbb{1}_{w_1>0}\mathbb{1}_{w_2>0}\mathbb{1}_{w_3>1} - 1.5 \cdot \mathbb{1}_{w_1>0}\mathbb{1}_{w_2>0}\mathbb{1}_{w_3\leq 1} \\
& -2.7 \cdot \mathbb{1}_{w_1>0}\mathbb{1}_{w_2\leq 0}\mathbb{1}_{w_2\leq -0.5}\mathbb{1}_{w_1>1}\mathbb{1}_{w_3>1.25} \\
& -0.5 \cdot \mathbb{1}_{w_1>0}\mathbb{1}_{w_2\leq}\mathbb{1}_{w_2\leq -0.5}\mathbb{1}_{w_1>1}\mathbb{1}_{w_3\leq 1.25} + 3.2 \cdot \mathbb{1}_{w_1>0}\mathbb{1}_{w_2\leq 0}\mathbb{1}_{w_2\leq -0.5}\mathbb{1}_{w_1\leq 1} \\
& +0.75 \cdot \mathbb{1}_{w_1>0}\mathbb{1}_{w_2\leq 0}\mathbb{1}_{w_2>-0.5} + 3 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3>0}\mathbb{1}_{w_2\leq -1}\mathbb{1}_{w_1\leq -1.3} \\
& +1.5 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3>0}\mathbb{1}_{w_2\leq -1}\mathbb{1}_{w_1>-1.3} - 2.3 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3>0}\mathbb{1}_{w_2>-1} \\
& +2.8 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3\leq -0.75} + 2 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3>-0.75}\mathbb{1}_{w_1\leq -0.5} \\
& -1.75 \cdot \mathbb{1}_{w_1\leq 0}\mathbb{1}_{w_3\leq 0}\mathbb{1}_{w_3>-0.75}\mathbb{1}_{w_1>-0.5}
\end{aligned}
$$

In the case of nonsmooth g, we consider $g(w) := 0.25 \cdot (1.3 \cdot w_1)^2$.

For the nonparametric covariable, we consider the following data generating mechanism. The matrix $\mathbf{W}_i \in \mathbb{R}^{n_i \times 3}$ contains the n_i observations of the ith experimental unit in its rows. We draw these n_i rows of $\mathbf{W}_i$ independently. That is, $(\mathbf{W}_i)_{k,\cdot} \sim \mathcal{N}_3(\mathbf{0}, \mathbb{1})$ for $i \in [N]$ and $k \in [n_i]$ with $(\mathbf{W}_i)_{k,\cdot} \perp\!\!\!\perp (\mathbf{W}_i)_{l,\cdot}$, $k \neq l, k, l \in [n_i]$ and $\mathbf{W}_i \perp\!\!\!\perp \mathbf{W}_j$, $i \neq j$, $i, j \in [N]$.

The linear covariable $\mathbf{X}_i$ is modeled with $\mathbf{X}_i = h(\mathbf{W}_i) + \boldsymbol{\varepsilon}_{\mathbf{X}_i}$, where its error term $\boldsymbol{\varepsilon}_{\mathbf{X}_i} \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbb{1})$ for $i \in [N]$ and $\boldsymbol{\varepsilon}_{\mathbf{X}_i} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{\mathbf{X}_j}$ for $i \neq j$, $i, j \in [N]$.

For $\beta_0 = 0.5$ and $\sigma_0 = 1$, the model of the response $\mathbf{Y}_i$ is $\mathbf{Y}_i = \mathbf{X}_i\beta_0 +$

$g(\mathbf{W}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$ with

$$\mathbf{Z}_i = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{n_i \times 3}, \quad \mathbf{b}_i = \begin{pmatrix} b_1^1 \\ b_2^1 \\ b^2 \end{pmatrix} \sim \mathcal{N}_3(\mathbf{0}, \operatorname{diag}(1.5^2, 1.8^2, 1.8^2)),$$

$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_0^2 \mathbb{1})$ for $i \in [N]$, and $\mathbf{b}_i \perp\!\!\!\perp \mathbf{b}_j$, $\mathbf{b}_i \perp\!\!\!\perp (\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j)$, and $\boldsymbol{\varepsilon}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_j$ for $i \neq j$, $i, j \in [N]$, where the first column of $\mathbf{Z}_i$ consists of $\lfloor 0.5 n_i \rfloor$ entries of 1's and $\lceil 0.5 n_i \rceil$ entries of 0's and correspondingly for the second column of $\mathbf{Z}_i$.

3.B | Assumptions and Additional Definitions

Recall the partially linear mixed-effects model

$$\mathbf{Y}_i = \mathbf{X}_i\beta_0 + g(\mathbf{W}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i \in [N]$$

for groups $i \in [N]$ as in (3.2). We consider N grouped observations $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)\}_{i \in [N]}$ from this model that satisfy Assumption 3.2.1. In each group $i \in [N]$, we observe n_i observations. We assume that these numbers are uniformly bounded by $n_{\max} < \infty$, that is, $n_i \leq n_{\max}$ for all $i \in [N]$. We denote the total number of observations of all groups by $N_T := \sum_{i=1}^N n_i$.

Let the number of sample splits $K \geq 2$ be a fixed integer independent of N. We assume that $N \geq K$ holds. Consider a partition $I_1, \ldots, I_K$ of $[N]$. For $k \in [K]$, we denote by $n_{T,k} := \sum_{i \in I_k} n_i$ the total number of observations of all groups i belonging to $I_k \subset [N]$. The sets $I_1, \ldots, I_K$ are assumed to be of approximately equal size in the sense that $K n_{T,k} = N_T + o(1)$ holds for all $k \in [K]$ as $N \to \infty$, which implies $\frac{N_T}{n_{T,k}} = O(1)$. Moreover, we assume that $\frac{|I_k|}{n_{T,k}} = O(1)$ holds for all $k \in [K]$.

For $k \in [K]$, denote by $\mathbf{S}_{I_k^c} := \{\mathbf{S}_i\}_{i \in I_k^c}$ the grouped observations from I_k^c. We denote the nuisance parameter estimator that is estimated with data from I_k^c by $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\mathbf{S}_{I_k^c})$.

Definition 3.B.1. *For $k \in [K]$, $\theta \in \Theta$, and $\eta \in \mathcal{T}$, where Θ and $\mathcal{T}$ are defined in Assumptions 3.B.3 and 3.B.4, respectively, we introduce the notation*

$$\mathbb{E}_{n_{T,k}}[\psi(\mathbf{S}; \theta, \eta)] := \frac{1}{n_{T,k}} \sum_{i \in I_k} \psi(\mathbf{S}_i; \theta, \eta).$$

Let $\{\delta_N\}_{N\geq K}$ and $\{\Delta_N\}_{N\geq K}$ be two sequences of non-negative numbers that converge to 0 as $N \to \infty$, where $\delta_N^2 \geq N^{-\frac{1}{2}}$ holds. We assume that $|I_k|^{-\frac{1}{2}+\frac{1}{p}} \log(|I_k|) \lesssim \delta_N$ holds for all $k \in [K]$, where p is specified in Assumption 3.B.2. Let $\{\mathcal{P}_N\}_{N\geq 1}$ be a sequence of sets of probability distributions P of the N grouped observations $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Y}_i)\}_{i\in[N]}$. We make the following additional assumptions.

Assumptions 3.B.2. *Let $p \geq 8$. For all N, all $i \in [N]$, all $P \in \mathcal{P}_N$, and all $k \in [K]$, we have the following.*

3.B.2.1 At the true θ_0 and the true η^0, the data $\{\mathbf{S}_i = (\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)\}_{i\in[N]}$ satisfies the identifiability condition

$$\mathbb{E}_P\left[\,\mathbb{E}_{n_{T,k}}[\psi(\mathbf{S};\theta_0,\eta^0)]\right] = \mathbf{0}.$$

3.B.2.2 There exists a finite real constant C_1 satisfying $\|\mathbf{X}_i\|_{P,p}+\|\mathbf{Y}_i\|_{P,p} \leq C_1$ for all $i \in [N]$.

3.B.2.3 The matrices $\mathbf{Z}_i$ assigning the random effects inside a group are fixed and bounded. In particular, there exists a finite real constant C_2 satisfying $\|\mathbf{Z}_i\| \leq C_2$ for all $i \in [N]$.

3.B.2.4 In absolute value, the smallest and largest singular values of the Jacobian matrix

$$J_0 := \partial_\theta\, \mathbb{E}_P\left[\,\mathbb{E}_{n_{T,k}}\left[\psi(\mathbf{S};\theta,\eta^0)\right]\right]\Big|_{\theta=\theta_0}$$

are bounded away from 0 by $c_1 > 0$ and are bounded away from $+\infty$ by $c_2 < \infty$.

3.B.2.5 For all $\theta \in \Theta$, we have the identification condition

$$\min\{\|J_0(\theta-\theta_0)\|, c_1\} \leq 2\Big\|\,\mathbb{E}_P\left[\,\mathbb{E}_{n_{T,k}}\left[\psi(\mathbf{S};\theta,\eta^0)\right]\right]\Big\|.$$

3.B.2.6 The matrix $\mathbb{E}_P[\mathbf{R}_{\mathbf{X}_i}^T(\mathbf{Z}_i\Sigma_0\mathbf{Z}_i + \sigma_0^2\mathbb{1}_{n_i})^{-1}\mathbf{R}_{\mathbf{X}_i}] \in \mathbb{R}^{d\times d}$ exists and is invertible for all $i \in [N]$. We assume that the same holds if θ_0 and η^0 are replaced by $\theta \in \Theta$ and $\eta \in \mathcal{T}$, respectively, with Θ as in Assumption 3.B.3 and $\mathcal{T}$ as in Assumption 3.B.4.

3.B.2.7 The symmetric matrix $\mathbb{E}_P[\mathbf{R}_{\mathbf{X}_i}^T(\mathbf{Z}_i\Sigma_0\mathbf{Z}_i + \sigma_0^2\mathbb{1}_{n_i})^{-1}\mathbf{R}_{\mathbf{X}_i}] \in \mathbb{R}^{d\times d}$ has singular values that are uniformly bounded away from 0 by $c_{\min} > 0$ for all $i \in [N]$.

3.B.2.8 There exists a symmetric positive-definite matrix $T_0 \in \mathbb{R}^{d\times d}$ *satisfying*

$$\overline{T}_N := \frac{1}{N_T}\sum_{i=1}^{N} \mathbb{E}_P\left[\mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{0,i}^{-1}\mathbf{R}_{\mathbf{X}_i}\right] = T_0 + o(1).$$

Assumption 3.B.2.1 ensures that β_0 is identifiable by our estimation method. Assumption 3.B.2.2 ensures that enough moments of $\mathbf{X}_i$ and $\mathbf{Y}_i$ exist. Assumption 3.B.2.4 and 3.B.2.5 are required to prove that θ_0 is consistently estimated in Lemma 3.C.7. The proof of this lemma uses a Taylor expansion. Assumption 3.B.2.6, 3.B.2.7, and 3.B.2.8 are required to make statements about the asymptotic variance-covariance matrix in the proof of Theorem 3.2.2.

The following Assumption 3.B.3 characterizes the set Θ to which θ_0 belongs and from which estimators of θ_0 are not too far away in the sense of Assumption 3.B.3.3.

Assumptions 3.B.3. *Consider the set*

$$\Theta := \left\{\theta = (\beta, \Sigma, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}^{q\times q} \times \mathbb{R} \colon \Sigma \in \mathbb{R}^{q\times q} \text{ symmetric positive definite, } \sigma > 0\right\}$$

of parameters. We make the following assumptions on Θ *and* $\hat{\theta}_k$ *for* $k \in [K]$.

3.B.3.1 The set Θ *is bounded and contains* θ_0 *and a ball of radius* $\max_{N\geq 1}\delta_N$ *around* θ_0.

3.B.3.2 There exists a finite real constant C_3 *such that we have* $\|(\mathbf{Z}_i\Sigma\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}\| \leq C_3$ *for all* $i \in [N]$ *and all* Σ *belonging to* Θ.

3.B.3.3 For all $k \in [K]$, *the estimator* $\hat{\theta}_k$ *belongs to* Θ *and satisfies the approximate solution property*

$$\left\|\mathbb{E}_{n_{T,k}}\left[\psi\big(\mathbf{S};\hat{\theta}_k,\hat{\eta}^{I_k^c}\big)\right]\right\| \leq \inf_{\theta\in\Theta}\left\|\mathbb{E}_{n_{T,k}}\left[\psi\big(\mathbf{S};\theta,\hat{\eta}^{I_k^c}\big)\right]\right\| + e_N$$

with the nuisance parameter estimator $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\boldsymbol{S}_{I_k^c})$, *where* $\{e_N\}_{N\geq K}$ *is a sequence of non-negative numbers satisfying* $e_N \lesssim \delta_N^2$.

The following Assumption 3.B.4 mainly characterizes the $N^{-1/2}$ product convergence rate of the machine learners that estimate the conditional expectations, which are nuisance functions.

Assumptions 3.B.4. *Consider the* $p \geq 8$ *from Assumption 3.B.2. For all* $N \geq K$ *and all* $P \in \mathcal{P}_N$, *consider a nuisance function realization set* $\mathcal{T}$ *such that the following conditions hold.*

3.B.4.1 The set $\mathcal{T}$ consists of P-integrable functions $\eta = (m_X, m_Y)$ whose pth moment exists, and it contains η^0. Furthermore, there exists a finite real constant C_4 such that

$$\begin{aligned}&\|\eta^0 - \eta\|_{P,p} \leq C_4, \quad \|\eta^0 - \eta\|_{P,2} \leq \delta_N^8,\\ &\|m_X^0(W) - m_X(W)\|_{P,2}\\ &\quad \cdot\big(\|m_Y^0(W) - m_Y(W)\|_{P,2} + \|m_X^0(W) - m_X(W)\|_{P,2}\big) \leq \delta_N N^{-\frac{1}{2}}\end{aligned}$$

hold for all elements η of $\mathcal{T}$.

3.B.4.2 For all $k \in [K]$, the nuisance parameter estimate $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\boldsymbol{S}_{I_k^c})$ satisfies

$$\begin{aligned}&\|\eta^0 - \hat{\eta}^{I_k^c}\|_{P,p} \leq C_4, \quad \|\eta^0 - \hat{\eta}^{I_k^c}\|_{P,2} \leq \delta_N^8,\\ &\|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\\ &\quad \cdot\big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \leq \delta_N N^{-\frac{1}{2}}\end{aligned}$$

with P-probability no less than $1 - \Delta_N$. Denote by $\mathcal{E}_N$ the event that $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\boldsymbol{S}_{I_k^c})$, $k \in [K]$ belong to $\mathcal{T}$, and assume this event holds with P-probability at least $1 - \Delta_N$.

3.B.4.3 For all $k \in [K]$, the parameter estimator $\hat{\theta}_k$ is P-integrable and its pth moment exists.

We suppose all assumptions presented in Section 3.B of the appendix hold throughout the remainder of the appendix.

3.C | Proof of Theorem 3.2.2

3.C.1 | Supplementary Lemmata

Lemma 3.C.1. *(Emmenegger and Bühlmann, 2021, Lemma G.7) Let $u \geq 1$. Consider a t-dimensional random variable A and an s-dimensional random variable B. Denote the joint law of A and B by P. Then, we have*

$$\|A - \mathbb{E}_P[A|B]\|_{P,u} \leq 2\|A\|_{P,u}.$$

Lemma 3.C.2. *(Emmenegger and Bühlmann, 2021, Lemma G.10) Consider a t_1-dimensional random variable A_1, a t_2-dimensional random variable A_2, and an s-dimensional random variable B. Denote the joint law of A_1, A_2, and B by P. Then, we have*

$$\big\|\,\mathbb{E}_P\big[(A_1 - \mathbb{E}_P[A_1|B])A_2^T\big]\big\|^2 \leq \|A_1\|_{P,2}^2\|A_2\|_{P,2}^2.$$

The following lemma, proved in Chernozhukov et al. (2018, Lemma 6.1) and Emmenegger and Bühlmann (2021, Lemma G.12), states that conditional convergence in probability implies unconditional convergence in probability.

Lemma 3.C.3. *(Chernozhukov et al. (2018, Lemma 6.1); Emmenegger and Bühlmann (2021, Lemma G.12)) Let $\{A_n\}_{n\geq 1}$ and $\{B_n\}_{n\geq 1}$ be sequences of random vectors, and let $u \geq 1$. Consider a deterministic sequence $\{\varepsilon_n\}_{n\geq 1}$ with $\varepsilon_n \to 0$ as $n \to \infty$ such that $\mathbb{E}[\|A_n\|^u | B_n] \leq \varepsilon_n^u$ holds. Then, we have $\|A_n\| = O_P(\varepsilon_n)$ unconditionally, meaning that that for any sequence $\{\ell_n\}_{n\geq 1}$ with $\ell_n \to \infty$ as $n \to \infty$, we have $P(\|A_n\| > \ell_n \varepsilon_n) \to 0$.*

3.C.2 | Representation of the Score Function ψ

Lemma 3.C.4. *Let $i \in [N]$, $\theta \in \Theta$, and $\eta \in \mathcal{T}$. Denote by $\mathbf{V}_i := \mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i}$. Furthermore, denote by ψ_β the coordinates of ψ that correspond to β, that is, $\psi_\beta(\mathbf{S}_i; \theta, \eta) = \nabla_\beta \ell_i(\theta, \eta)$. We have*

$$\psi_\beta(\mathbf{S}_i; \theta, \eta) = \frac{1}{\sigma^2}\big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)^T \mathbf{V}_i^{-1}\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big).$$

Proof. The statement follows from the definition of ψ. □

Lemma 3.C.5. *Let $i \in [N]$, $\theta \in \Theta$, and $\eta \in \mathcal{T}$. Denote by $\mathbf{V}_i := \mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i}$. Furthermore, denote by ψ_{σ^2} the coordinates of ψ that correspond to σ^2, that is, $\psi_{\sigma^2}(\mathbf{S}_i; \theta, \eta) = \nabla_{\sigma^2} \ell_i(\theta, \eta)$. We have*

$$\begin{aligned}\psi_{\sigma^2}(\mathbf{S}_i; \theta, \eta) = &-\tfrac{n_i}{2\sigma^2} + \tfrac{1}{2(\sigma^2)^2}\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big)^T \mathbf{V}_i^{-1} \\ &\cdot\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big).\end{aligned}$$

Proof. The statement follows from the definition of ψ. □

Lemma 3.C.6. *Let $i \in [N]$, $\theta \in \Theta$, $\eta \in \mathcal{T}$. Denote by $\mathbf{V}_i := \mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i}$. Furthermore, let indices $\kappa, \iota \in [q]$, and denote by $\psi_{\Sigma_{\kappa,\iota}}$ the coordinates of ψ that correspond to $\Sigma_{\kappa,\iota}$, that is, $\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i; \theta, \eta) = \nabla_{\Sigma_{\kappa,\iota}} \ell_i(\theta, \eta)$. We have*

$$\begin{aligned}&\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i; \theta, \eta) \\ = \; &-\tfrac{1}{2}\textstyle\sum_{t,u=1}^{n_i} (\mathbf{V}_i^{-1})_{t,u} (\mathbf{Z}_i)_{t,\kappa} (\mathbf{Z}_i^T)_{\iota,u} \\ &+\tfrac{1}{2\sigma^2}\textstyle\sum_{t,u=1}^{n_i} \Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big)_t \\ &\cdot\Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta\Big)_u \big(\mathbf{V}_i^{-1} (\mathbf{Z}_i)_{\cdot,\kappa} (\mathbf{Z}_i^T)_{\iota,\cdot} \mathbf{V}_i^{-1}\big)_{t,u}.\end{aligned}$$

Proof. Let a vector $x \in \mathbb{R}^{n_i}$. We have

$$\begin{aligned} & \tfrac{\partial}{\partial \Sigma_{\kappa,\iota}} \Big(x^T \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} x \Big) \\ = \; & \Sigma_{t,u=1}^{n_i} \tfrac{\partial}{\partial (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})_{t,u}^{-1}} \Big(x^T \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} x \Big) \cdot \tfrac{\partial (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})_{t,u}^{-1}}{\partial \Sigma_{\kappa,\iota}}. \end{aligned}$$

For some nonrandom matrix $D \in \mathbb{R}^{n_i \times n_i}$, we have

$$\frac{\partial}{\partial D_{t,u}} x^T D x = \frac{\partial}{\partial D_{t,u}} \sum_{r,s=1}^{n_i} x_r D_{r,s} x_s = x_t x_u.$$

Furthermore, we have

$$\frac{\partial}{\partial \Sigma_{\kappa,\iota}} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} = - \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} \Big(\frac{\partial}{\partial \Sigma_{\kappa,\iota}} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big) \Big) \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1}$$

by Petersen and Pedersen (2012, Equation (59)), and we have

$$\Big(\frac{\partial}{\partial \Sigma_{\kappa,\iota}} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big) \Big)_{t,u} = \frac{\partial}{\partial \Sigma_{\kappa,\iota}} \sum_{r,s=1}^{n_i} (\mathbf{Z}_i)_{t,r} \Sigma_{r,s} (\mathbf{Z}_i^T)_{s,u} = (\mathbf{Z}_i)_{t,\kappa} (\mathbf{Z}_i^T)_{\iota,u},$$

and consequently

$$\frac{\partial}{\partial \Sigma_{\kappa,\iota}} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big) = (\mathbf{Z}_i)_{\cdot,\kappa} (\mathbf{Z}_i^T)_{\iota,\cdot},$$

which leads to

$$\frac{\partial}{\partial \Sigma_{\kappa,\iota}} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} = - \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} (\mathbf{Z}_i)_{\cdot,\kappa} (\mathbf{Z}_i^T)_{\iota,\cdot} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1}.$$

Therefore, we have

$$\begin{aligned} & \tfrac{\partial}{\partial \Sigma_{\kappa,\iota}} \Big(x^T \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} x \Big) \\ = \; & - \Sigma_{t,u=1}^{n_i} x_t x_u \cdot \Big(\big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} (\mathbf{Z}_i)_{\cdot,\kappa} (\mathbf{Z}_i^T)_{\iota,\cdot} \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big)^{-1} \Big)_{t,u}. \end{aligned} \tag{3.9}$$

Moreover, we have

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_{\kappa,\iota}} \log \Big(\det \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big) \Big) \\ = & \sum_{t,u=1}^{n_i} \frac{\partial}{\partial (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})_{t,u}} \log \Big(\det \big(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \big) \Big) \cdot \frac{\partial (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})_{t,u}}{\partial \Sigma_{\kappa,\iota}} \end{aligned}$$

$$= \sum_{t,u=1}^{n_i} \left(\left(\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i} \right)^{-1} \right)_{t,u} (\mathbf{Z}_i)_{t,\kappa} (\mathbf{Z}_i^T)_{\iota,u} \tag{3.10}$$

by Petersen and Pedersen (2012, Equation (57)). We replace x in (3.9) by $\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta$ and combine (3.9) and (3.10) to conclude the proof. □

3.C.3 | Consistency

This section establishes that all $\hat{\theta}_k$, $k \in [K]$ are consistent. In particular, this implies that $\hat{\theta}$ is consistent.

Let $P \in \mathcal{P}_N$.

Lemma 3.C.7. *Let $k \in [K]$. We have $\|\hat{\theta}_k - \theta_0\| \leq \delta_N^2$ with P-probability $1 - o(1)$.*

Proof of Lemma 3.C.7. We have

$$\begin{aligned} & \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \eta^0) \right] \right] \\ = \; & \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \eta^0) \right] - \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \hat{\eta}^{I_k^c}) \right] \right] \\ & + \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \hat{\eta}^{I_k^c}) \right] \right] - \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \hat{\eta}^{I_k^c}) \right] + \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \hat{\eta}^{I_k^c}) \right]. \end{aligned} \tag{3.11}$$

Due to the approximate solution property in Assumption 3.B.3.3, the identifiability condition in Assumption 3.B.2.1, and the triangle inequality, we have

$$\begin{aligned} & \leq \left\| \begin{matrix} \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \hat{\theta}_k, \hat{\eta}^{I_k^c}) \right] \\ \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta_0, \hat{\eta}^{I_k^c}) \right] \end{matrix} \right\| + e_N \\ & \leq \left\| \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta_0, \hat{\eta}^{I_k^c}) \right] - \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta_0, \hat{\eta}^{I_k^c}) \right] \right] \right\| \\ & \quad + \left\| \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta_0, \hat{\eta}^{I_k^c}) \right] \right] - \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta_0, \eta^0) \right] \right] \right\| + e_N. \end{aligned} \tag{3.12}$$

Let us introduce

$$\mathcal{I}_1 := \sup_{\substack{\theta \in \Theta, \\ \eta \in \mathcal{T}}} \left\| \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta, \eta) \right] \right] - \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta, \eta^0) \right] \right] \right\| \tag{3.13}$$

and

$$\mathcal{I}_2 := \sup_{\theta \in \Theta} \left\| \mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta, \hat{\eta}^{I_k^c}) \right] - \mathbb{E}_P \left[\mathbb{E}_{n_{T,k}} \left[\psi(\mathbf{S}; \theta, \hat{\eta}^{I_k^c}) \right] \right] \right\|. \tag{3.14}$$

Due to (3.11) and (3.12), we infer, with P-probability $1-o(1)$,

$$\Big\| \mathbb{E}_P \Big[\mathbb{E}_{n_{T,k}} \big[\psi\big(\mathbf{S}; \hat{\theta}_k, \eta^0\big) \big] \Big] \Big\| \leq e_N + 2\mathcal{I}_1 + 2\mathcal{I}_2$$

because the event $\mathcal{E}_N$ that $\hat{\eta}^{I_k^c}$ belongs to $\mathcal{T}$ holds with P-probability $1-o(1)$ by Assumption 3.B.4.2. By Lemma 3.C.8, we have $\mathcal{I}_1 \lesssim \delta_N^2$. By Lemma 3.C.10, we have $\mathcal{I}_2 \lesssim N^{-\frac{1}{2}}$ with P-probability $1-o(1)$. Recall that we have $\delta_N^2 \geq N^{-\frac{1}{2}}$ and $e_N \lesssim \delta_N^2$. With P-probability $1-o(1)$, we therefore have

$$\min\{\|J_0(\hat{\theta}_k - \theta_0)\|, c_1\} \leq 2 \Big\| \mathbb{E}_P \Big[\mathbb{E}_{n_{T,k}} \big[\psi(\mathbf{S}; \hat{\theta}_k, \eta^0) \big] \Big] \Big\| \lesssim \delta_N^2$$

due to Assumption 3.B.2.5. We infer our claim because the singular values of J_0 are bounded away from 0 by Assumption 3.B.2.4. □

Lemma 3.C.8. *Consider*

$$\mathcal{I}_1 = \sup_{\substack{\theta \in \Theta, \\ \eta \in \mathcal{T}}} \Big\| \mathbb{E}_P \Big[\mathbb{E}_{n_{T,k}} \big[\psi(\mathbf{S}; \theta, \eta) \big] \Big] - \mathbb{E}_P \Big[\mathbb{E}_{n_{T,k}} \big[\psi\big(\mathbf{S}; \theta, \eta^0\big) \big] \Big] \Big\|$$

as in (3.13). *We have* $\mathcal{I}_1 \lesssim \delta_N^2$.

Proof. Let indices $i \in [N]$ and $\kappa, \iota \in [q]$, let $\theta \in \Theta$, and let $\eta \in \mathcal{T}$. Furthermore, let $\psi_\beta(\mathbf{S}_i; \theta, \eta) := \nabla_\beta \ell_i(\theta, \eta)$, let $\psi_{\sigma^2}(\mathbf{S}_i; \theta, \eta) := \nabla_{\sigma^2} \ell_i(\theta, \eta)$, and let $\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i; \theta, \eta) := \nabla_{\Sigma_{\kappa,\iota}} \ell_i(\theta, \eta)$. Denote by $\mathbf{V}_i := \mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i}$. We have

$$\begin{aligned}
&\psi_\beta(\mathbf{S}_i; \theta, \eta) - \psi_\beta\big(\mathbf{S}_i; \theta, \eta^0\big) \\
= \; & \tfrac{1}{\sigma^2} \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)^T \mathbf{V}_i^{-1} \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big) \\
& + \tfrac{1}{\sigma^2} \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)^T \mathbf{V}_i^{-1} \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta \Big) \\
& + \tfrac{1}{\sigma^2} \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)^T \mathbf{V}_i^{-1} \\
& \quad \cdot \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big),
\end{aligned} \tag{3.15}$$

we have

$$\begin{aligned}
&\psi_{\sigma^2}(\mathbf{S}_i; \theta, \eta) - \psi_{\sigma^2}\big(\mathbf{S}_i; \theta, \eta^0\big) \\
=& 2 \cdot \frac{1}{2(\sigma^2)^2} \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta \Big)^T \mathbf{V}_i^{-1} \\
& \quad \cdot \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big) \\
& + \frac{1}{2(\sigma^2)^2} \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big)^T \mathbf{V}_i^{-1}
\end{aligned}$$

$$\cdot \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big), \tag{3.16}$$

and we have

$$\begin{aligned}
&\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i;\theta,\eta) - \psi_{\Sigma_{\kappa,\iota}}\big(\mathbf{S}_i;\theta,\eta^0\big) \\
=\ & \tfrac{1}{2\sigma^2} \textstyle\sum_{t,u=1}^{n_i} \big(\mathbf{V}_i^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_i^{-1}\big)_{t,u} \\
&\cdot \Bigg(\Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big)_t \\
&\qquad \cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta \Big)_u \\
&\quad + \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta \Big)_t \\
&\qquad \cdot \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big)_u \\
&\quad + \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big)_t \\
&\qquad \cdot \Big(m_Y^0(\mathbf{W}_i) - m_Y(\mathbf{W}_i) - \big(m_X^0(\mathbf{W}_i) - m_X(\mathbf{W}_i)\big)\beta \Big)_u \Bigg).
\end{aligned} \tag{3.17}$$

Up to constants depending on the diameter of Θ, the L^1-norms of all terms (3.15)–(3.17) are bounded by δ_N due to Hölder's inequality because we have $n_i \leq n_{\max}$, $\|\mathbf{X}_i - m_X^0(\mathbf{W}_i)\|_{P,2} \leq \|\mathbf{X}_i\|_{P,2}$ by Lemma 3.C.2 and similarly for $\mathbf{Y}_i$, $\|\mathbf{X}_i\|_{P,2}$ and $\|\mathbf{Y}_i\|_{P,2}$ are bounded by Assumption 3.B.2.2 and Hölder's inequality, $\mathbf{Z}_i$ is bounded by Assumption 3.B.2.3, $\mathbf{V}_i^{-1} = (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ is bounded by Assumption 3.B.3.2, $\|\eta^0 - \eta\|_{P,2} \leq \delta_N^8 \leq \delta_N^2$ holds by Assumption 3.B.4.1 for N large enough, and Θ is bounded by Assumption 3.B.3.1. Therefore, we infer the claim. □

Lemma 3.C.9. *Let $\eta \in \mathcal{T}$, and consider the function class defined by $\mathcal{F}_\eta := \{\psi_j(\cdot;\theta,\eta) \colon j \in [d+1+q^2], \theta \in \Theta\}$. Let $i \in [N]$ and $\theta_1, \theta_2 \in \Theta$. Then, there exists a function $h \in L^2$ such that for all $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}_\eta$, we have*

$$|f_{\theta_1}(\cdot) - f_{\theta_2}(\cdot)| \leq h(\cdot)\|\theta_1 - \theta_2\|.$$

Proof. Let $i \in [N]$, and consider the grouped observations $\mathbf{S}_i$ of group i. Independently of i, the number of observations n_i from this group is bounded by $n_{\max} < \infty$.

Let $\eta = (m_X, m_Y) \in \mathcal{T}$, and let $\theta_1, \theta_2 \in \Theta$. Denote by $\mathbf{V}_{i,1} := \mathbf{Z}_i \Sigma_1 \mathbf{Z}_i^T + \mathbb{1}_{n_i}$, and denote by $\mathbf{V}_{i,2} := \mathbf{Z}_i \Sigma_2 \mathbf{Z}_i^T + \mathbb{1}_{n_i}$. Moreover, denote by $\mathbf{R}_{\mathbf{X}_i,\eta} := \mathbf{X}_i - m_X(\mathbf{W}_i)$ and by $\mathbf{R}_{\mathbf{Y}_i,\eta} := \mathbf{Y}_i - m_Y(\mathbf{W}_i)$. Furthermore, consider indices $\kappa, \iota \in [q]$, and let $\psi_\beta(\mathbf{S}_i;\theta,\eta) := \nabla_\beta \ell_i(\theta,\eta)$, let $\psi_{\sigma^2}(\mathbf{S}_i;\theta,\eta) := \nabla_{\sigma^2}\ell_i(\theta,\eta)$,

and let $\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i;\theta,\eta) := \nabla_{\Sigma_{\kappa,\iota}}\ell_i(\theta,\eta)$. Observe that

$$\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1} = \mathbf{V}_{i,1}^{-1}\big(\mathbf{V}_{i,2} - \mathbf{V}_{i,1}\big)\mathbf{V}_{i,2}^{-1} \tag{3.18}$$

and

$$\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} = \big(\sigma_1^2\big)^{-1}\big(\sigma_2^2 - \sigma_1^2\big)\big(\sigma_2^2\big)^{-1} \tag{3.19}$$

hold. Thus, we have

$$\begin{aligned}
&\psi_\beta(\mathbf{S}_i;\theta_1,\eta) - \psi_\beta\big(\mathbf{S}_i;\theta_2,\eta\big)\\
=\;& \Big(\tfrac{1}{\sigma_1^2} - \tfrac{1}{\sigma_2^2}\Big)\mathbf{R}_{\mathbf{X}_i,\eta}^T\mathbf{V}_{i,1}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}^T\beta_1\big) + \tfrac{1}{\sigma_2^2}\mathbf{R}_{\mathbf{X}_i,\eta}^T\mathbf{V}_{i,1}^{-1}\big(\mathbf{V}_{i,2} - \mathbf{V}_{i,1}\big)\mathbf{V}_{i,2}^{-1}\mathbf{R}_{\mathbf{Y}_i,\eta}\\
&-\tfrac{1}{\sigma_2^2}\mathbf{R}_{\mathbf{X}_i,\eta}^T\Big(\mathbf{V}_{i,1}^{-1}\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_1-\beta_2) + \mathbf{V}_{i,1}^{-1}\big(\mathbf{V}_{i,2} - \mathbf{V}_{i,1}\big)\mathbf{V}_{i,2}^{-1}\mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\Big),
\end{aligned}$$

and

$$\begin{aligned}
&\psi_{\sigma^2}(\mathbf{S}_i;\theta_1,\eta) - \psi_{\sigma^2}\big(\mathbf{S}_i;\theta_2,\eta\big)\\
=\;& \tfrac{n_i}{2}\Big(\tfrac{1}{\sigma_2^2} - \tfrac{1}{\sigma_1^2}\Big)\\
&+\tfrac{1}{2(\sigma_1^2)^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)^T\mathbf{V}_{i,1}^{-1}\big(\mathbf{V}_{i,2} - \mathbf{V}_{i,1}\big)\mathbf{V}_{i,2}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)\\
&+\tfrac{1}{2}\Big(\tfrac{1}{\sigma_1^2} - \tfrac{1}{\sigma_2^2}\Big)\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)^T\mathbf{V}_{i,2}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)\\
&+\tfrac{2}{2(\sigma_2^2)^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)\mathbf{V}_{i,2}^{-1}\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2-\beta_1)\\
&+\tfrac{1}{2(\sigma_2^2)^2}(\beta_2-\beta_1)^T\mathbf{R}_{\mathbf{X}_i,\eta}^T\mathbf{V}_{i,2}^{-1}\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2-\beta_1),
\end{aligned}$$

and

$$\begin{aligned}
&\psi_{\Sigma_{\kappa,\iota}}(\mathbf{S}_i;\theta_1,\eta) - \psi_{\Sigma_{\kappa,\iota}}\big(\mathbf{S}_i;\theta_2,\eta\big)\\
=\;& -\tfrac{1}{2}\textstyle\sum_{t,u=1}^{n_i}\Big(\mathbf{V}_{i,1}^{-1}\big(\mathbf{V}_{i,2} - \mathbf{V}_{i,1}\big)\mathbf{V}_{i,2}^{-1}\Big)_{t,u}(\mathbf{Z}_i)_{t,\kappa}(\mathbf{Z}_i^T)_{\iota,u}\\
&+\tfrac{1}{2}\textstyle\sum_{t,u=1}^{n_i}\Big(\tfrac{1}{\sigma_1^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_t\\
&\qquad\cdot\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_u\big(\mathbf{V}_{i,1}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,1}^{-1}\big)_{t,u}\\
&\qquad-\tfrac{1}{\sigma_2^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_u\big(\mathbf{V}_{i,2}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,2}^{-1}\big)_{t,u}\Big),
\end{aligned}$$

where for $t,u\in[n_i]$, we have

$$\begin{aligned}
&\tfrac{1}{\sigma_1^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_u\big(\mathbf{V}_{i,1}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,1}^{-1}\big)_{t,u}\\
=\;& \Big(\tfrac{1}{\sigma_1^2} - \tfrac{1}{\sigma_2^2}\Big)\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_u\big(\mathbf{V}_{i,1}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,1}^{-1}\big)_{t,u}\\
&+\tfrac{1}{\sigma_2^2}\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_u\big(\mathbf{V}_{i,1}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,1}^{-1}\big)_{t,u}
\end{aligned}$$

and

$$
\begin{aligned}
&\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_1\big)_u\big(\mathbf{V}_{i,1}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,1}^{-1}\big)_{t,u}\\
&\quad-\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_u\big(\mathbf{V}_{i,2}^{-1}(\mathbf{Z}_i)_{\cdot,\kappa}(\mathbf{Z}_i^T)_{\iota,\cdot}\mathbf{V}_{i,2}^{-1}\big)_{t,u}\\
=\;&\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_u\\
&\quad\cdot\Big(\big(\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1}\big)_{t,\cdot}(\mathbf{Z}_i)_{\cdot,\kappa}\big(\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1}\big)_{u,\cdot}(\mathbf{Z}_i)_{\cdot,\iota}\\
&\quad+\big(\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1}\big)_{t,\cdot}(\mathbf{Z}_i)_{\cdot,\kappa}\big(\mathbf{V}_{i,2}^{-1}\big)_{u,\cdot}(\mathbf{Z}_i)_{\cdot,\iota}\\
&\quad+\big(\mathbf{V}_{i,2}^{-1}\big)_{t,\cdot}(\mathbf{Z}_i)_{\cdot,\kappa}\big(\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1}\big)_{u,\cdot}(\mathbf{Z}_i)_{\cdot,\iota}\Big)\\
&\quad+\Big(\big(\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2 - \beta_1)\big)_t\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_u\\
&\quad+\big(\mathbf{R}_{\mathbf{Y}_i,\eta} - \mathbf{R}_{\mathbf{X}_i,\eta}\beta_2\big)_t\big(\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2 - \beta_1)\big)_u\\
&\quad+\big(\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2 - \beta_1)\big)_t\big(\mathbf{R}_{\mathbf{X}_i,\eta}(\beta_2 - \beta_1)\big)_u\Big)\big(\mathbf{V}_{i,1}^{-1}\big)_{t,\cdot}(\mathbf{Z}_i)_{\cdot,\kappa}\big(\mathbf{V}_{i,1}^{-1}\big)_{u,\cdot}(\mathbf{Z}_i)_{\cdot,\iota}.
\end{aligned}
$$

Due to (3.18), the terms $\mathbf{V}_{i,1}^{-1} - \mathbf{V}_{i,2}^{-1}$ can be represented in terms of $\mathbf{V}_{i,2} - \mathbf{V}_{i,1}$. Due to (3.19), the terms $(\sigma_1^2)^{-1} - (\sigma_2^2)^{-1}$ can be represented in terms of $\sigma_2^2 - \sigma_1^2$. Recall that $n_i \leq n_{\max}$, $\|\mathbf{X}_i - m_X^0(\mathbf{W}_i)\|_{P,2} \leq \|\mathbf{X}_i\|_{P,2}$ by Lemma 3.C.2 and similarly for $\mathbf{Y}_i$, $\|\mathbf{X}_i\|_{P,2}$ and $\|\mathbf{Y}_i\|_{P,2}$ are bounded by Assumption 3.B.2.2 and Hölder's inequality, $\mathbf{Z}_i$ is bounded by Assumption 3.B.2.3, $\mathbf{V}_{i,1}^{-1} = (\mathbf{Z}_i\Sigma_1\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ and $\mathbf{V}_{i,2}^{-1} = (\mathbf{Z}_i\Sigma_2\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ are bounded by Assumption 3.B.3.2, m_X and m_Y are square integrable by Assumption 3.B.4.1, and Θ is bounded by Assumption 3.B.3.1. Therefore, we infer the claim. □

Lemma 3.C.10. *Consider*

$$
\mathcal{I}_2 = \sup_{\theta\in\Theta}\left\|\mathbb{E}_{n_{T,k}}\left[\psi\big(\mathbf{S};\theta,\hat{\eta}^{I_k^c}\big)\right] - \mathbb{E}_P\left[\mathbb{E}_{n_{T,k}}\left[\psi\big(\mathbf{S};\theta,\hat{\eta}^{I_k^c}\big)\right]\right]\right\|
$$

as in (3.14). *We have* $\mathcal{I}_2 \lesssim N^{-\frac{1}{2}}$ *with* P*-probability* $1 - o(1)$.

Proof. The proof of the statement follows from Lemma 3.C.11. □

A version of the following lemma with not only independent but also identically distributed random variables is presented in Chernozhukov et al. (2018, Lemma 6.2) and in Chernozhukov et al. (2014, Theorem 5.1, Corollary 5.1). However, as we subsequently show, their results can be generalized to only requiring independence.

Lemma 3.C.11. *(Maximal Inequality: Chernozhukov et al. (2018, Lemma 6.2); Chernozhukov et al. (2014, Theorem 5.1, Corollary 5.1)) Let* $\eta \in \mathcal{T}$, *and consider the function class* $\mathcal{F}_\eta := \{\psi_j(\cdot;\theta,\eta)\colon j \in [d + 1 + q^2], \theta \in \Theta\}$. *Suppose that* $F_\eta \geq \sup_{f\in\mathcal{F}_\eta}|f|$ *is a measurable envelope for* $\mathcal{F}_\eta$ *with* $\|F_\eta\|_{P,p} < \infty$. *Let* $k \in [K]$, *and let* $M := \max_{i\in I_k} F_\eta(\mathbf{S}_i)$. *Let* $\tau^2 > 0$ *be*

a positive constant satisfying $\sup_{f\in\mathcal{F}_\eta}\|f\|_{P,2}^2 \le \tau^2 \le \|F_\eta\|_{P,2}^2 < \infty$*, where we write* $\|\varphi\|_{P,2}^2 = \frac{1}{|I_k|}\sum_{i\in I_k}\mathbb{E}_P[\varphi^2(\mathbf{S}_i)]$ *for functions* φ*. Suppose there exist constants* $a \ge e$ *and* $v \ge 1$ *such that for all* $0 < \varepsilon \le 1$*,*

$$\log \sup_Q N(\varepsilon\|F_\eta\|_{Q,2}, \mathcal{F}_\eta, \|\cdot\|_{Q,2}) \le v\log(a/\varepsilon) \tag{3.20}$$

holds, where Q *runs over the class* $\{|I_k|^{-1}\sum_{i\in I_k} Q_i \colon Q_i$ *a probability measure*$\}$ *of measures. Consider the empirical process*

$$\mathbb{G}_{P,I_k}[\psi(\mathbf{S})] := \frac{1}{\sqrt{|I_k|}}\sum_{i\in I_k}\big(\psi(\mathbf{S}_i) - \mathbb{E}_P[\psi(\mathbf{S}_i)]\big).$$

Then, we have

$$\mathbb{E}_P[\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_\eta}] \le C\cdot\left(\sqrt{v\tau^2\log\big(a\|F_\eta\|_{P,2}\tau^{-1}\big)} + \frac{v\|M\|_{P,2}}{\sqrt{|I_k|}}\log\big(a\|F_\eta\|_{P,2}\tau^{-1}\big)\right), \tag{3.21}$$

where C *is an absolute constant. Moreover, for every* $t \ge 1$*, with probability* $> 1 - t^{-\frac{p}{2}}$*, we have*

$$\begin{aligned}\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_\eta} &\le (1+\alpha)\,\mathbb{E}_P[\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_\eta}] \\ &\quad + C(p)\Big((\tau + |I_k|^{-\frac{1}{2}}\|M\|_{P,p})\sqrt{t} + \alpha^{-1}|I_k|^{-\frac{1}{2}}\|M\|_{P,2}t\Big)\end{aligned} \tag{3.22}$$

for all $\alpha > 0$*, where* $C(p) > 0$ *is a constant depending only on* p*. In particular, setting* $a \ge |I_k|$ *and* $t = \log(|I_k|)$*, with probability* $> 1 - c\cdot(\log(|I_k|))^{-1}$ *for some constant* c*, we have*

$$\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_\eta} \le C(p,c)\left(\tau\sqrt{v\log\big(a\|F_\eta\|_{P,2}\tau^{-1}\big)} + \frac{v\|M\|_{P,2}}{\sqrt{|I_k|}}\log\big(a\|F_\eta\|_{P,2}\tau^{-1}\big)\right),$$

where $\|M\|_{P,p} \le |I_k|^{\frac{1}{p}}\|F_\eta\|_{P,p}$ *and* $C(p,c) > 0$ *is a constant depending only on* p *and* c*.*

Proof. Observe that an envelope F_η as described in the lemma exists due to Lemma 3.C.9. Consequently, statement (3.20) holds with $a = \operatorname{diam}(\Theta)$ due to van der Vaart (1998, Example 19.7) and due to Lemma 3.C.9. Liu et al. (2021) proceed similarly to establish a similar claim. The proof of Chernozhukov et al. (2014, Corollary 5.1) can be adapted to verify statement (3.21), and the proof of Chernozhukov et al. (2014, Theorem 5.1) can be adapted to show statement (3.22). Adaptations are required because these two results are stated for independent and identically distributed data. Our grouped data $\{\mathbf{S}_i\}_{i\in[N]}$ is groupwise independent, but not identically distributed because a different

number of observations may be available for the different groups $i \in [N]$. Subsequently, we describe these adaptations.

The proof of Chernozhukov et al. (2014, Theorem 5.1) is based on Boucheron et al. (2005, Theorem 12). The latter result is an inequality for functions of independent random variables and does not require identically distributed variables. Thus, statement (3.22) is established in our setting.

Also the proof of Chernozhukov et al. (2014, Theorem 5.2) only requires independent but not necessarily identically distributed random variables. Hence, the Corollary 5.1 of Theorem 5.2 in Chernozhukov et al. (2014) remains to hold in our setting, and thus statement (3.21) is established as well. □

3.C.4 | Asymptotic Distribution of the Fixed-Effects Estimator

Proof of Theorem 3.2.2. Fix a sequence $\{P_N\}_{N\geq 1}$ of probability measures such that $P_N \in \mathcal{P}_N$ for all $N \geq 1$. Because this sequence is chosen arbitrarily, it suffices to show that (3.8) holds along $\{P_N\}_{N\geq 1}$ to infer that it holds uniformly over $P \in \mathcal{P}_N$.

Recall the notations $\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i} = \mathbf{X}_i - \hat{m}^{I_k^c}_X(\mathbf{W}_i)$ and $\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} = \mathbf{Y}_i - \hat{m}^{I_k^c}_Y(\mathbf{W}_i)$ for $i \in [N]$. Observe that the estimator $\hat{\beta}$ in (3.7) can alternatively be represented by

$$\hat{\beta} = \frac{1}{K}\sum_{k=1}^{K}\left(\arg\min_{\beta}\frac{1}{n_{T,k}}\sum_{i\in I_k}\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta\big)^T\hat{\mathbf{V}}^{-1}_{i,k}\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta\big)\right)$$

for $\hat{\mathbf{V}}_{i,k} := \mathbf{Z}_i\hat{\Sigma}_k\mathbf{Z}_i^T + \mathbb{1}_{n_i}$ because the Gaussian likelihood decouples. In particular, $\hat{\beta}$ has a generalized least squares representation. Observe furthermore that we have

$$\begin{aligned}&\sqrt{N_T}(\hat{\beta} - \beta_0)\\ &= \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\left(\frac{1}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i})^T\hat{\mathbf{V}}^{-1}_{i,k}\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\right)^{-1}\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i})^T\hat{\mathbf{V}}^{-1}_{i,k}\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta_0\big).\end{aligned} \tag{3.23}$$

Let $k \in [K]$. We have

$$\begin{aligned}&\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i})^T\hat{\mathbf{V}}^{-1}_{i,k}\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta_0\big)\\ =\ &\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i})^T\mathbf{V}^{-1}_{i,0}\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta_0\big)\\ &+\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i})^T\big(\hat{\mathbf{V}}^{-1}_{i,k} - \mathbf{V}^{-1}_{i,0}\big)\big(\widehat{\mathbf{R}}^{I_k}_{\mathbf{Y}_i} - \widehat{\mathbf{R}}^{I_k}_{\mathbf{X}_i}\beta_0\big).\end{aligned} \tag{3.24}$$

We analyze the two terms in the above decomposition (3.24) individually. We start with the second term. For $i \in [N]$, $\eta \in \mathcal{T}$, and Σ from Θ, define the

function

$$\begin{aligned}&\varphi(\mathbf{S}_i;\Sigma,\eta)\\&:=\big(\mathbf{X}_i-m_X(\mathbf{W}_i)\big)^T(\mathbf{Z}_i\Sigma\mathbf{Z}_i^T+\mathbb{1}_{n_i})^{-1}\Big(\mathbf{Y}_i-m_Y(\mathbf{W}_i)-\big(\mathbf{X}_i-m_X(\mathbf{W}_i)\big)\beta_0\Big).\end{aligned}\tag{3.25}$$

We have

$$\begin{aligned}&\tfrac{\sqrt{N_T}}{n_{T,k}}\textstyle\sum_{i\in I_k}(\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T\big(\hat{\mathbf{V}}_{i,k}^{-1}-\mathbf{V}_{i,0}^{-1}\big)\big(\widehat{\mathbf{R}}_{\mathbf{Y}_i}^{I_k}-\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k}\beta_0\big)\\=\;&\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})\big]\\=\;&\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big]\\&-\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big].\end{aligned}\tag{3.26}$$

Next, we analyze the two terms in (3.26). The second term is of order

$$\big\|\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big]\big\|=o_{P_N}(1)\tag{3.27}$$

by Lemma 3.C.12. The first term in (3.26) is bounded by

$$\begin{aligned}&\big\|\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big]\big\|\\\le\;&\sup_{\|\Sigma-\Sigma_0\|\le\delta_N}\big\|\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\Sigma,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big]\big\|.\end{aligned}$$

with P_N-probability $1-o(1)$ due to Lemma 3.C.7 because we have $\delta_N^2\le\delta_N$ for N large enough. Let Σ be from Θ with $\|\Sigma-\Sigma_0\|\le\delta_N$. With P_N-probability $1-o(1)$, we have

$$\sqrt{N_T}\,\mathbb{E}_{n_{T,k}}\big[\varphi(\mathbf{S};\Sigma,\hat{\eta}^{I_k^c})-\varphi(\mathbf{S};\Sigma_0,\eta^0)\big]\lesssim\delta_N\tag{3.28}$$

by Lemma 3.C.15. Consequently, the second term in (3.24) is of order $o_{P_N}(1)$ due to (3.26), (3.27), and (3.28). Subsequently, we analyze the first term in (3.24). By Lemma 3.C.12, we have

$$\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T\mathbf{V}_{i,0}^{-1}\big(\widehat{\mathbf{R}}_{\mathbf{Y}_i}^{I_k}-\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k}\beta_0\big)=\frac{\sqrt{N_T}}{n_{T,k}}\sum_{i\in I_k}\mathbf{R}_{\mathbf{X}_i}^T\mathbf{V}_{i,0}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i}-\mathbf{R}_{\mathbf{X}_i}\beta_0\big)+o_{P_N}(1).$$

Denote by

$$T_{N,i}:=\mathbb{E}_{P_N}\left[\mathbf{R}_{\mathbf{X}_i}^T\mathbf{V}_{i,0}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i}-\mathbf{R}_{\mathbf{X}_i}\beta_0\big)\Big(\mathbf{R}_{\mathbf{X}_i}^T\mathbf{V}_{i,0}^{-1}\big(\mathbf{R}_{\mathbf{Y}_i}-\mathbf{R}_{\mathbf{X}_i}\beta_0\big)\Big)^T\right].$$

We have

$$T_{N,i}=\mathbb{E}_{P_N}\left[\mathbf{R}_{\mathbf{X}_i}^T\mathbf{V}_{i,0}^{-1}\mathbf{R}_{\mathbf{X}_i}\right]\tag{3.29}$$

due to Assumption 3.2.1.4. Furthermore, recall $\overline{T}_N=\frac{1}{N_T}\Sigma_{i=1}^N T_{N,i}$ from

Assumption 3.B.2.8. Due to Assumption 3.B.2.7, the singular values of the matrices $T_{N,i}$, $i \in [N]$ are uniformly bounded away from 0 by $c_{\min} > 0$. Thus, the smallest eigenvalue ν_N^2 of $\overline{T}_N$ satisfies

$$\nu_N^2 \geq \frac{1}{N_T} \sum_{i=1}^{N} \lambda_{\min}\Big(T_{N,i}\Big) \geq \frac{1}{n_{\max}} c_{\min} > 0 \tag{3.30}$$

because we have $N_T \leq N n_{\max}$ with $n_{\max} < \infty$. Next, we verify the Lindeberg condition. Due to the Cauchy-Schwarz inequality, Markov's inequality, Hölder's inequality, and (3.30), we have

$$\begin{aligned}
& \tfrac{1}{N_T \nu_N^2} \textstyle\sum_{i=1}^{N} \mathbb{E}_{P_N}\left[\left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \right\|^2 \mathbb{1}_{\left\{ \left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} (\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0) \right\|^2 \geq \varepsilon N_T \nu_N^2 \right\}} \right] \\
\leq \; & \tfrac{1}{N_T \nu_N^2} \textstyle\sum_{i=1}^{N} \left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \right\|_{P_N,4}^2 \\
& \cdot \sqrt{P_N\Big(\left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} (\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0) \right\|^2 \geq \varepsilon N_T \nu_N^2 \Big)} \\
\leq \; & \tfrac{1}{N_T \nu_N^2} \textstyle\sum_{i=1}^{N} \left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \right\|_{P_N,4}^2 \left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \right\|_{P_N,2} \\
& \cdot \sqrt{\tfrac{1}{\varepsilon N_T \nu_N^2}} \\
\leq \; & \tfrac{1}{N_T \nu_N^2} \textstyle\sum_{i=1}^{N} \left\| \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \right\|_{P_N,4}^3 \sqrt{\tfrac{1}{\varepsilon N_T \nu_N^2}} \\
\lesssim \; & \sqrt{\tfrac{1}{\varepsilon N_T}} \overset{N \to \infty}{\longrightarrow} 0
\end{aligned}$$

for $\varepsilon > 0$ by Assumptions 3.B.2.2, 3.B.3.1, 3.B.3.2, and Lemma 3.C.1. Consequently, we have

$$(\overline{T}_N)^{-\frac{1}{2}} \frac{1}{\sqrt{N_T}} \sum_{i=1}^{N} \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \xrightarrow{\mathcal{L}} \mathcal{N}_d(\mathbf{0}, \mathbb{1})$$

by Hansen (2017, Theorem 6.9.2). Thus, we infer

$$\begin{aligned}
& (\overline{T}_N)^{-\frac{1}{2}} \sqrt{N_T} \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \tfrac{1}{n_{T,k}} \sum_{i \in I_k} \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) \\
= \; & (\overline{T}_N)^{-\frac{1}{2}} \tfrac{1}{\sqrt{N_T}} \textstyle\sum_{i=1}^{N} \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) + o_{P_N}(1)
\end{aligned}$$

due to $n_{T,k} = \frac{N_T}{K} = o(1)$.

Finally, the term $\frac{1}{n_{T,k}} \sum_{i \in I_k} (\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T \hat{\mathbf{V}}_{i,k}^{-1} \widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k}$ in (3.23) equals $T_0 + o_{P_N}(1)$ due to Lemma 3.C.18. Therefore, we have

$$\sqrt{N_T} T_0^{\frac{1}{2}} (\hat{\beta} - \beta_0) = (\overline{T}_N)^{-\frac{1}{2}} \frac{1}{\sqrt{N_T}} \sum_{i=1}^{N} \mathbf{R}_{\mathbf{X}_i}^T \mathbf{V}_{i,0}^{-1} \big(\mathbf{R}_{\mathbf{Y}_i} - \mathbf{R}_{\mathbf{X}_i} \beta_0 \big) + o_{P_N}(1) \xrightarrow{\mathcal{L}} \mathcal{N}_d(\mathbf{0}, \mathbb{1}).$$

□

Lemma 3.C.12. *Let $k \in [K]$. For $i \in [N]$ and $\eta \in \mathcal{T}$, consider the*

function

$$\begin{aligned}&\varphi(\mathbf{S}_i; \Sigma_0, \eta)\\ &= \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \Big(\mathbf{Y}_i - m_Y(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)\beta_0\Big)\end{aligned}$$

as in (3.25)*, but where we consider* Σ_0 *instead of general* Σ *from* Θ*. We have*

$$\left\| \frac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \varphi(\mathbf{S}_i; \Sigma_0, \hat{\eta}^{I_k^c}) - \frac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \varphi(\mathbf{S}_i; \Sigma_0, \eta^0) \right\| = O_P(\delta_N).$$

Proof of Lemma 3.C.12. A similar proof that is modified from Chernozhukov et al. (2018) is presented in Emmenegger and Bühlmann (2021, Lemma G.16). For notational simplicity, we omit the argument Σ_0 in φ and write $\varphi(\mathbf{S}_i; \eta)$ instead of $\varphi(\mathbf{S}_i; \Sigma_0, \eta)$. By the triangle inequality, we have

$$\begin{aligned}&\left\| \tfrac{\sqrt{N_T}}{n_{T,k}} \textstyle\sum_{i \in I_k} \varphi(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \tfrac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \varphi(\mathbf{S}_i; \eta^0) \right\| \\ = &\left\| \begin{array}{l} \tfrac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \big(\varphi(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \int \varphi(\mathbf{s}_i; \hat{\eta}^{I_k^c}) \, \mathrm{d}P(\mathbf{s}_i)\big) \\ \quad - \tfrac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \big(\varphi(\mathbf{S}_i; \eta^0) - \int \varphi(\mathbf{s}_i; \eta^0) \, \mathrm{d}P(\mathbf{s}_i)\big) \\ \quad + \sqrt{N_T} \tfrac{1}{n_{T,k}} \sum_{i \in I_k} \int \big(\varphi(\mathbf{s}_i; \hat{\eta}^{I_k^c}) - \varphi(\mathbf{s}_i; \eta^0)\big) \, \mathrm{d}P(\mathbf{s}_i) \end{array} \right\| \\ \leq &\ \mathcal{I}_1 + \sqrt{N_T} \mathcal{I}_2,\end{aligned}$$

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

$$\begin{aligned}M := &\tfrac{\sqrt{N_T}}{n_{T,k}} \textstyle\sum_{i \in I_k} \Big(\varphi(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \int \varphi(\mathbf{s}_i; \hat{\eta}^{I_k^c}) \, \mathrm{d}P(\mathbf{s}_i)\Big) - \\ &\tfrac{\sqrt{N_T}}{n_{T,k}} \textstyle\sum_{i \in I_k} \Big(\varphi(\mathbf{S}_i; \eta^0) - \int \varphi(\mathbf{s}_i; \eta^0) \, \mathrm{d}P(\mathbf{s}_i)\Big),\end{aligned}$$

and where

$$\mathcal{I}_2 := \left\| \frac{1}{n_{T,k}} \sum_{i \in I_k} \int \big(\varphi(\mathbf{s}_i; \hat{\eta}^{I_k^c}) - \varphi(\mathbf{s}_i; \eta^0)\big) \, \mathrm{d}P(\mathbf{s}_i) \right\|.$$

Subsequently, we bound the two terms $\mathcal{I}_1$ and $\mathcal{I}_2$ individually. First, we bound $\mathcal{I}_1$. Because the dimensions d of β_0 and q of the random effects model are fixed, it is sufficient to bound one entry of the d-dimensional column vector M. Let

$t \in [d]$. On the event $\mathcal{E}_N$ that holds with P-probability $1 - o(1)$, we have

$$\begin{aligned}
& \mathbb{E}_P\left[\|M_t\|^2 \middle| \mathbf{S}_{I_k^c}\right] \\
= \; & \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\left|\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_i; \eta^0)\right|^2 \middle| \mathbf{S}_{I_k^c}\right] \\
& + \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}_P\left[\left(\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_i; \eta^0)\right)\right. \\
& \qquad \left. \cdot \left(\varphi_t(\mathbf{S}_j; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_j; \eta^0)\right) \middle| \mathbf{S}_{I_k^c}\right] \\
& - \tfrac{2N_T}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_i; \eta^0) \middle| \mathbf{S}_{I_k^c}\right] \\
& \qquad \cdot \textstyle\sum_{j \in I_k} \mathbb{E}_P\left[\varphi_t(\mathbf{S}_j; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_j; \eta^0) \middle| \mathbf{S}_{I_k^c}\right] \\
& + \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(S_i; \eta^0) \middle| \mathbf{S}_{I_k^c}\right]^2 \\
& + \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}_P\left[\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_i; \eta^0) \middle| \mathbf{S}_{I_k^c}\right] \\
& \qquad \cdot \mathbb{E}_P\left[\varphi_t(\mathbf{S}_j; \hat{\eta}^{I_k^c}) - \varphi_t(\mathbf{S}_j; \eta^0) \middle| \mathbf{S}_{I_k^c}\right] \\
\leq \; & \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\left|\varphi_t(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \varphi_t(S_i; \eta^0)\right|^2 \middle| \mathbf{S}_{I_k^c}\right] \\
\leq \; & \sup_{\eta \in \mathcal{T}} \tfrac{N_T}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[\|\varphi(\mathbf{S}_i; \eta) - \varphi(\mathbf{S}_i; \eta^0)\|^2\right]
\end{aligned} \tag{3.31}$$

because $\mathbf{S}_i$ and $\mathbf{S}_j$ are independent for $i \neq j$. Due to Lemma 3.C.13, we have $\mathbb{E}_P[\mathcal{I}_1^2 | \mathbf{S}_{I_k^c}] \lesssim \delta_N^4 \leq \delta^2$ for N large enough because $\frac{N_T}{n_{T,k}}$ is of order $O(1)$ by assumption. Thus, we infer $\mathcal{I}_1 = O_P(\delta_N)$ by Lemma 4.I.12. Subsequently, we bound $\mathcal{I}_2$. Let $i \in I_k$. For $r \in [0, 1]$, we introduce the function

$$f_k(r) := \frac{1}{n_{T,k}} \sum_{i \in I_k} \left(\mathbb{E}_P\left[\varphi\big(\mathbf{S}_i; \eta^0 + r(\hat{\eta}^{I_k^c} - \eta^0)\big) \middle| \mathbf{S}_{I_k^c}\right] - \mathbb{E}_P[\varphi(\mathbf{S}_i; \eta^0)] \right).$$

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

$$f_k(0) = \frac{1}{n_{T,k}} \sum_{i \in I_k} \left(\mathbb{E}_P\left[\varphi(\mathbf{S}_i; \eta^0) \middle| \mathbf{S}_{I_k^c}\right] - \mathbb{E}_P[\varphi(\mathbf{S}_i; \eta^0)] \right) = \mathbf{0}.$$

Furthermore, the score φ satisfies the Neyman orthogonality property $f_k'(0) = \mathbf{0}$ on the event $\mathcal{E}_N$ that holds with P-probability $1 - o(1)$ because we have for all $i \in I_k$ and $\eta \in \mathcal{T}$ that

$$\begin{aligned}
& \frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\left[\varphi\big(\mathbf{S}_i; \eta^0 + r(\eta - \eta^0)\big)\right] \\
= & \frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\left[\Big(\mathbf{X}_i - m_X^0(\mathbf{W}_i) - r\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\Big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}\right.
\end{aligned}$$

$$
\begin{aligned}
&\cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - r\big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i)\big) \\
&\qquad - \Big(\mathbf{X}_i - m_X^0(\mathbf{W}_i) - r\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\Big)\beta_0 \Big) \Big] \\
=& \, \mathbb{E}_P \Big[- \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \\
&\qquad \cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta_0 \Big) \\
&\quad - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)(\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \\
&\qquad \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0 \Big) \Big] \\
=& \mathbf{0}
\end{aligned} \tag{3.32}
$$

holds because we can apply the tower property to condition on $\mathbf{W}_i$ inside the above expectations, and because m_X^0 and m_Y^0 are the true conditional expectations. Moreover, we have

$$
\begin{aligned}
& \tfrac{\partial^2}{\partial r^2} \, \mathbb{E}_P \left[\varphi\big(\mathbf{S}_i; \eta^0 + r(\eta - \eta^0)\big) \right] \\
= \; & 2 \, \mathbb{E}_P \Big[\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \\
& \qquad \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0 \Big) \Big]
\end{aligned}
$$

for all $i \in I_k$. On the event $\mathcal{E}_N$ that holds with P-probability $1 - o(1)$, we have

$$
\|f_k''(\tilde{r})\| \leq \sup_{r \in (0,1)} \|f_k''(r)\| \lesssim \delta_N N^{-\frac{1}{2}}
$$

by Lemma 3.C.14. Therefore, we conclude

$$
\left\| \frac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \varphi(\mathbf{S}_i; \hat{\eta}^{I_k^c}) - \frac{\sqrt{N_T}}{n_{T,k}} \sum_{i \in I_k} \varphi(\mathbf{S}_i; \eta^0) \right\| \leq \mathcal{I}_1 + \sqrt{N_T} \mathcal{I}_2 = O_P(\delta_N).
$$

□

Lemma 3.C.13. *We have*

$$
\sup_{\eta \in \mathcal{T}} \frac{1}{n_{T,k}} \sum_{i \in I_k} \mathbb{E}_P \left[\|\varphi(\mathbf{S}_i; \Sigma_0, \eta) - \varphi(\mathbf{S}_i; \Sigma_0, \eta^0)\|^2 \right] \lesssim \delta_N^4.
$$

Proof of Lemma 3.C.13. A similar proof that is modified from Chernozhukov et al. (2018) is presented in Emmenegger and Bühlmann (2021, Lemma G.15, Lemma G.16). For notational simplicity, we omit the argument Σ_0 in φ and write $\varphi(\mathbf{S}_i; \eta)$ instead of $\varphi(\mathbf{S}_i; \Sigma_0, \eta)$. Recall the notation $\mathbf{V}_{0,i} = \mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i}$

for $i \in [N]$. Because we have $\sup_{i\in[N]}\|\mathbf{V}_{0,i}^{-1}\| \leq C_3$ by Assumption 3.B.3.2, we have

$$\frac{1}{n_{T,k}} \sum_{i\in I_k} \mathbb{E}_P\left[\left\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\right\|\right] \lesssim \delta_N^8 \tag{3.33}$$

by the triangle inequality, Hölder's inequality, and because we have for all $i \in I_k$ that $n_i \leq n_{\max}$, $\|\mathbf{X}_i - m_X^0(\mathbf{W}_i)\|_{P,2} \leq \|\mathbf{X}_i\|_{P,2}$ by Lemma 3.C.2 and similarly for $\mathbf{Y}_i$, $\|\mathbf{X}_i\|_{P,2}$ and $\|\mathbf{Y}_i\|_{P,2}$ are bounded by Assumption 3.B.2.2 and Hölder's inequality, $(\mathbf{Z}_i\Sigma_0\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ is bounded by Assumption 3.B.3.2, and $\|\eta^0 - \eta\|_{P,2} \leq \delta_N^8$ holds by Assumption 3.B.4.1.

Furthermore, we have

$$\begin{aligned} & \mathbb{E}_P\left[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|^2\right] \\ \leq\ & \mathbb{E}_P[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|] \\ & + \mathbb{E}_P\left[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|^2 \mathbb{1}_{\|\varphi(\mathbf{S}_i;\eta)-\varphi(\mathbf{S}_i;\eta^0)\|\geq 1}\right], \end{aligned} \tag{3.34}$$

and we have

$$\begin{aligned} & \mathbb{E}_P\left[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|^2 \mathbb{1}_{\|\varphi(\mathbf{S}_i;\eta)-\varphi(\mathbf{S}_i;\eta^0)\|\geq 1}\right] \\ \leq\ & \left\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\right\|_{P,4}^2 \sqrt{P(\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\| \geq 1)} \end{aligned} \tag{3.35}$$

by Hölder's inequality. Observe that the term

$$\left\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\right\|_{P,4}^2 \tag{3.36}$$

is upper bounded by the triangle inequality, Hölder's inequality, because we have $n_i \leq n_{\max}$, $\|\mathbf{X}_i - m_X^0(\mathbf{W}_i)\|_{P,p} \lesssim \|\mathbf{X}_i\|_{P,p}$ by Lemma 3.C.1 and similarly for $\mathbf{Y}_i$, $\|\mathbf{X}_i\|_{P,p}$ and $\|\mathbf{Y}_i\|_{P,p}$ are bounded by Assumption 3.B.2.2, $(\mathbf{Z}_i\Sigma_0\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ is bounded by Assumption 3.B.3.2, and $\|\eta^0 - \eta\|_{P,p}$ is upper bounded by Assumption 3.B.4.1. By Markov's inequality, we furthermore have

$$P(\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\| \geq 1) \leq \mathbb{E}_P[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|] \leq \delta_N^8 \tag{3.37}$$

due to (3.33). Therefore, we have

$$\sup_{\eta\in\mathcal{T}} \frac{1}{n_{T,k}} \sum_{i\in I_k} \mathbb{E}_P\left[\|\varphi(\mathbf{S}_i;\eta) - \varphi(\mathbf{S}_i;\eta^0)\|^2\right] \lesssim \delta_N^8 + \delta_N^4 \lesssim \delta_N^4$$

for N large enough due to (3.33)–(4.29). □

Lemma 3.C.14. *Let $\eta \in \mathcal{T}$, and let $i \in [N]$. We have*

$$\mathbb{E}_P \Big[\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T \mathbf{V}_{0,i}^{-1} \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0 \Big) \Big] \lesssim \delta_N N^{-\frac{1}{2}}.$$

Proof of Lemma 3.C.14. The claim follows by applying Hölder's inequality and the Cauchy-Schwarz inequality because $\sup_{i\in[N]} \|\mathbf{V}_{0,i}^{-1}\|$ is upper bounded by Assumption 3.B.3.2, Θ is bounded, and

$$\|m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\|_{P,2} \cdot \big(\|m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i)\|_{P,2} + \|m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\|_{P,2} \big) \leq \delta_N N^{-\frac{1}{2}}$$

holds by Assumption 3.B.4.1. □

Lemma 3.C.15. *Let Σ from Θ with $\|\theta - \theta_0\| \leq \delta_N^2$. With P-probability $1 - o(1)$, we have*

$$\sqrt{N_T}\, \mathbb{E}_{n_{T,k}} \big[\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0) \big] \lesssim \delta_N.$$

Proof of Lemma 3.C.15. Observe that we have

$$\begin{aligned} & \sqrt{N_T}\, \mathbb{E}_{n_{T,k}} \big[\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0) \big] \\ = \; & \sqrt{\tfrac{N_T}{n_{T,k}}} \sqrt{\tfrac{|I_k|}{n_{T,k}}} \mathbb{G}_{P,I_k} \big[\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0) \big] \\ & + \sqrt{N_T}\, \mathbb{E}_P \big[\mathbb{E}_{n_{T,k}} [\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0)] | \mathbf{S}_{I_k^c} \big], \end{aligned}$$

where the second summand is bounded by δ_N due to Lemma 3.C.17, and where we recall the empirical process notation

$$\mathbb{G}_{P,I_k}[\varphi(\mathbf{S})] = \frac{1}{\sqrt{|I_k|}} \sum_{i \in I_k} \Big(\varphi(\mathbf{S}_i) - \int \varphi(\mathbf{s}_i)\, \mathrm{d}P(\mathbf{s}_i) \Big)$$

for some function φ. Consider the function class

$$\mathcal{F}_2 := \big\{ \varphi_j(\cdot; \Sigma, \hat{\eta}^{I_k^c}) - \varphi_j(\cdot; \Sigma_0, \eta^0) \colon j \in [d], \|\Sigma - \Sigma_0\| \leq \delta_N^2 \big\}.$$

We have $\sqrt{\frac{N_T}{n_{T,k}}} \sqrt{\frac{|I_k|}{n_{T,k}}} = O(1)$ by assumption. Therefore, it suffices to bound

$$\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_2} = \sup_{f \in \mathcal{F}_2} |\mathbb{G}_{P,I_k}[f]|.$$

To bound this term, we apply Lemma 3.C.11 conditional on $\mathbf{S}_{I_k^c}$ to the empirical process $\{\mathbb{G}_{P,I_k}[f] \colon f \in \mathcal{F}_2\}$ with the envelope $F_2 := F_{\hat{\eta}^{I_k^c}} + F_{\eta^0}$ and $\tau = C r'_{N,k}$

for a sufficiently large constant C, where $r'_{N,k}$ is defined by

$$r'_{N,k} := \sup_{\substack{\eta\in\mathcal{T},\\ \|\Sigma-\Sigma_0\|\le\delta_N^2}} \left\| \frac{1}{|I_k|}\sum_{i\in I_k} \varphi(\mathbf{S}_i;\Sigma,\eta) - \varphi(\mathbf{S}_i;\Sigma_0,\eta^0)\right\|_{P,2} \tag{3.38}$$

and satisfies $\sup_{f\in\mathcal{F}_2}\|f\|_{P,2} \lesssim r'_{N,k}$ with P-probability $1-o(1)$. The estimated nuisance parameter $\hat{\eta}^{I_k^c}$ can be treated as fixed if we condition on $\mathbf{S}_{I_k^c}$. Thus, with P-probability $1-o(1)$, we have

$$\sup_{f\in\mathcal{F}_2} |\mathbb{G}_{P,I_k}[f]| \lesssim r'_{N,k}\sqrt{\log\left(\frac{1}{r'_{N,k}}\right)} + |I_k|^{-\frac{1}{2}+\frac{1}{p}}\log(|I_k|) \tag{3.39}$$

because $\|F_2\|_{P,p} = \|F_{\hat{\eta}^{I_k^c}} + F_{\eta^0}\|_{P,p}$ is finite by the triangle inequality and Lemma 3.C.9, because $\mathcal{F}_2 \subset \mathcal{F}_{\hat{\eta}^{I_k^c}} - \mathcal{F}_{\eta^0}$, and because the uniform covering entropy satisfies

$$\begin{aligned}
&\log\sup_Q N\big(\varepsilon\|F_{\hat{\eta}^{I_k^c}} + F_{\eta^0}\|_{Q,2}, \mathcal{F}_{\hat{\eta}^{I_k^c}} - \mathcal{F}_{\eta^0}, \|\cdot\|_{Q,2}\big)\\
\le\ &\log\sup_Q N\big(\tfrac{\varepsilon}{2}\|F_{\hat{\eta}^{I_k^c}}\|_{Q,2}, \mathcal{F}_{\hat{\eta}^{I_k^c}}, \|\cdot\|_{Q,2}\big) + \log\sup_Q N\big(\tfrac{\varepsilon}{2}\|F_{\eta^0}\|_{Q,2}, \mathcal{F}_{\eta^0}, \|\cdot\|_{Q,2}\big)\\
\le\ &2v\log\left(\frac{2a}{\epsilon}\right)
\end{aligned}$$

for all $0<\varepsilon\le 1$ due to Andrews (1986, Proof of Theorem 3) as presented in Chernozhukov et al. (2018). We have $r'_{N,k} \le C\delta_N^2$ for some constant C due to Lemma 3.C.16. For N large enough, we have $r'_{N,k} < 1$. The function $\alpha\colon (0,1) \ni x \mapsto x\sqrt{\log(x^{-1})} \in \mathbb{R}$ is non-negative, increasing for x small enough, and satisfies $\lim_{x\to 0^+} x\sqrt{\log(x^{-1})} = 0$. Thus, we have $\alpha(r'_{N,k}) = o(1)$ and $\alpha(r'_{N,k}) \le \alpha(C\delta_N^2)$ for N large enough. Moreover, we have $\alpha(x) \lesssim \sqrt{x}$ for $x\in(0,1)$, so that we infer $\alpha(r'_{N,k}) \lesssim \delta_N$. Because we assumed $|I_k|^{-\frac{1}{2}+\frac{1}{p}}\log(|I_k|) \lesssim \delta_N$, we have $\|\mathbb{G}_{P,I_k}\|_{\mathcal{F}_2} \lesssim \delta_N$ with P-probability $1-o(1)$ as claimed due to (3.39). □

Lemma 3.C.16. *Let $k\in K$. Recall*

$$r'_{N,k} = \sup_{\substack{\eta\in\mathcal{T},\\ \|\Sigma-\Sigma_0\|\le\delta_N^2}} \left\| \frac{1}{|I_k|}\sum_{i\in I_k} \varphi(\mathbf{S}_i;\Sigma,\eta) - \varphi(\mathbf{S}_i;\Sigma_0,\eta^0)\right\|_{P,2}$$

from (3.38). *We have $r'_{N,k} \lesssim \delta_N^2$.*

Proof of Lemma 3.C.16. Let $\eta\in\mathcal{T}$, Σ from Θ with $\|\Sigma-\Sigma_0\| \le \delta_N^2$, and

$i \in [N]$. We have

$$\begin{aligned} & \varphi(\mathbf{S}_i; \Sigma, \eta) - \varphi(\mathbf{S}_i, \Sigma_0, \eta^0) \\ = \; & \varphi(\mathbf{S}_i; \Sigma, \eta) - \varphi(\mathbf{S}_i; \Sigma, \eta^0) + \varphi(\mathbf{S}_i; \Sigma, \eta^0) - \varphi(\mathbf{S}_i; \Sigma_0, \eta^0). \end{aligned}$$

Let $t \in [d]$. We have

$$\begin{aligned} & \|\mathbb{E}_{n_{T,k}}[\varphi_t(\mathbf{S}; \Sigma, \eta) - \varphi_t(\mathbf{S}; \Sigma, \eta^0)]\|_{P,2}^2 \\ = \; & \tfrac{1}{n_{T,k}^2} \textstyle\sum_{i \in I_k} \mathbb{E}_P\left[(\varphi_t(\mathbf{S}_i; \Sigma, \eta) - \varphi_t(\mathbf{S}_i; \Sigma_0, \eta^0))^2\right] \\ & + \tfrac{1}{n_{T,k}^2} \textstyle\sum_{i,j \in I_k, i \neq j} \mathbb{E}_P\left[\varphi_t(\mathbf{S}_i; \Sigma, \eta) - \varphi_t(\mathbf{S}_i; \Sigma_0, \eta^0)\right] \\ & \cdot \mathbb{E}_P\left[\varphi_t(\mathbf{S}_j; \Sigma, \eta) - \varphi_t(\mathbf{S}_j; \Sigma_0, \eta^0)\right] \\ \lesssim \; & \delta_N^4 \end{aligned}$$

due to $\mathbf{S}_i \perp\!\!\!\perp \mathbf{S}_j$ for $i \neq j$ and similar arguments as presented in the proof of Lemma 3.C.13. Furthermore, we have

$$\|\mathbb{E}_{n_{T,k}}[\varphi(\mathbf{S}_i; \Sigma, \eta^0) - \varphi(\mathbf{S}_i; \Sigma_0, \eta^0)]\|_{P,2} \lesssim \delta_N^2$$

due to the Cauchy-Schwarz inequality, $\|\Sigma - \Sigma_0\| \leq \delta_N^2$, because we have $n_i \leq n_{\max}$, $\|\mathbf{X}_i - m_X^0(\mathbf{W}_i)\|_{P,4} \lesssim \|\mathbf{X}_i\|_{P,4}$ by Lemma 3.C.1 and similarly for $\mathbf{Y}_i$, $\|\mathbf{X}_i\|_{P,4}$ and $\|\mathbf{Y}_i\|_{P,4}$ are bounded by Assumption 3.B.2.2 and Hölder's inequality, $\mathbf{Z}_i$ is bounded by Assumption 3.B.2.3, $\mathbf{V}_i^{-1} = (\mathbf{Z}_i \Sigma \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}$ is bounded by Assumption 3.B.3.2, $\|\eta^0 - \eta\|_{P,p} \leq C_4$ holds by Assumption 3.B.4.1 for N large enough, and Θ is bounded by Assumption 3.B.3.1. Consequently, we have $r'_{N,k} \lesssim \delta_N^2$ due to the triangle inequality. □

Lemma 3.C.17. *Let $k \in [K]$. For Σ belonging to Θ, with P-probability $1 - o(1)$, we have*

$$\left\|\sqrt{N_T}\, \mathbb{E}_P\left[\, \mathbb{E}_{n_{T,k}}[\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0)] \middle| \boldsymbol{S}_{I_k^c}\right]\right\| \lesssim \delta_N.$$

Proof of Lemma 3.C.17. With P-probability $1 - o(1)$, the machine learning estimator $\hat{\eta}^{I_k^c}$ belongs to the nuisance realization set $\mathcal{T}$ due to Assumption 3.B.4.2. Thus, it suffices to show that the claim holds uniformly over $\eta \in \mathcal{T}$. Consider $\eta \in \mathcal{T}$ and Σ belonging to Θ. For $r \in [0, 1]$, consider the function

$$f_k(r) := \mathbb{E}_P\left[\varphi\big(\mathbf{S}_i; \Sigma_0 + r(\Sigma - \Sigma_0), \eta^0 + r(\hat{\eta}^{I_k^c} - \eta^0)\big) \middle| \mathbf{S}_{I_k^c}\right] - \mathbb{E}_P[\varphi(\mathbf{S}_i; \Sigma_0, \eta^0)].$$

We apply a Taylor expansion to this function and obtain

$$\begin{aligned} & \sqrt{N_T}\, \mathbb{E}_P\left[\, \mathbb{E}_{n_{T,k}}[\varphi(\mathbf{S}; \Sigma, \hat{\eta}^{I_k^c}) - \varphi(\mathbf{S}; \Sigma_0, \eta^0)] \middle| \mathbf{S}_{I_k^c}\right] \\ = \; & \sqrt{N_T} f_k(1) \\ = \; & \sqrt{N_T}\big(f_k(0) + f_k'(0) + \tfrac{1}{2} f_k''(\tilde{r})\big) \end{aligned}$$

for some $\tilde{r} \in (0,1)$. We have $f_k(0) = \mathbf{0}$. Next, we verify the Neyman orthogonality property $f_k'(0) = \mathbf{0}$ and the second-order condition $f_k''(r) \lesssim \delta_N N^{-\frac{1}{2}}$ uniformly over $r \in (0,1)$, which will conclude the proof. We have

$$
\begin{aligned}
& \tfrac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\left[\varphi\big(\mathbf{S}_i; \Sigma_0 + r(\Sigma - \Sigma_0), \eta^0 + r(\eta - \eta^0)\big)\right] \\
= \; & \tfrac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}_P\Bigg[\Big(\mathbf{X}_i - m_X^0(\mathbf{W}_i) - r\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\Big)^T \\
& \qquad \cdot (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i} + r\mathbf{Z}_i(\Sigma - \Sigma_0)\mathbf{Z}_i^T)^{-1} \\
& \qquad \cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - r\big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i)\big) \\
& \qquad\qquad - \Big(\mathbf{X}_i - m_X^0(\mathbf{W}_i) - r\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\Big)\beta_0\Big)\Bigg] \\
= \; & \mathbb{E}_P\Big[- \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \\
& \qquad\qquad \cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta_0\Big) \\
& \qquad + \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)^T \Big(\tfrac{\partial}{\partial r}\Big|_{r=0} (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i} + r\mathbf{Z}_i(\Sigma - \Sigma_0)\mathbf{Z}_i^T)^{-1}\Big) \\
& \qquad\qquad \cdot \Big(\mathbf{Y}_i - m_Y^0(\mathbf{W}_i) - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)\beta_0\Big) \\
& \qquad - \big(\mathbf{X}_i - m_X^0(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1} \\
& \qquad\qquad \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0\Big)\Big) \\
= \; & \mathbf{0},
\end{aligned}
\tag{3.40}
$$

where we apply the tower property to condition on $\mathbf{W}_i$ inside the above expectation, Assumption 3.2.1.4, and that m_X^0 and m_Y^0 are the true conditional expectations. Thus, we have $f_k'(0) = \mathbf{0}$. Furthermore, we have

$$
\begin{aligned}
& \frac{\partial^2}{\partial r^2} \mathbb{E}_P\left[\varphi\big(\mathbf{S}_i; \Sigma_0 + r(\Sigma - \Sigma_0), \eta^0 + r(\eta - \eta^0)\big)\right] \\
= & \, \mathbb{E}_P\Big[2\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T (\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i} + r\mathbf{Z}_i(\Sigma - \Sigma_0)\mathbf{Z}_i^T)^{-1} \\
& \qquad \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0\Big)\Big] \\
& + 4r \, \mathbb{E}_P\Big[\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T \\
& \qquad \cdot \Big(\frac{\partial}{\partial r}(\mathbf{Z}_i \Sigma_0 \mathbf{Z}_i^T + \mathbb{1}_{n_i} + r\mathbf{Z}_i(\Sigma - \Sigma_0)\mathbf{Z}_i^T)^{-1}\Big) \\
& \qquad \cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0\Big)\Big] \\
& + r^2 \, \mathbb{E}_P\Big[\big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)^T
\end{aligned}
$$

$$\cdot \left(\frac{\partial^2}{\partial r^2}(\mathbf{Z}_i\Sigma_0\mathbf{Z}_i^T + \mathbb{1}_{n_i} + r\mathbf{Z}_i(\Sigma-\Sigma_0)\mathbf{Z}_i^T)^{-1}\right)$$
$$\cdot \Big(m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \big(m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i)\big)\beta_0\Big)\Big],$$

where we apply the tower property to condition on $\mathbf{W}_i$ inside the above expectation, Assumption 3.2.1.4, and that m_X^0 and m_Y^0 are the true conditional expectations. All the above summands are bounded by $\delta_N N^{-\frac{1}{2}}$ in the L^1-norm due to Hölder's inequality and Assumptions 3.B.2.2, 3.B.3.1, 3.B.3.2, and 3.B.4.1 because for $\mathbf{A}_i := m_X(\mathbf{W}_i) - m_X^0(\mathbf{W}_i) \in \mathbb{R}^{n_i\times d}$, $\mathbf{B}_i := m_Y(\mathbf{W}_i) - m_Y^0(\mathbf{W}_i) - \mathbf{A}_i\beta_0 \in \mathbb{R}^{n_i}$, and a nonrandom matrix $\mathbf{D}_i \in \mathbb{R}^{n_i\times n_i}$ with bounded entries, we have for $j \in [d]$ that

$$\begin{aligned}\|(\mathbf{A}_i^T)_{j,\cdot}\mathbf{D}_i\mathbf{B}_i\|_{P,1} &= \Big\|\textstyle\sum_{\kappa,\iota=1}^{n_i}(\mathbf{A}_i^T)_{j,\kappa}(\mathbf{D}_i)_{\kappa,\iota}(\mathbf{B}_i)_\iota\Big\|_{P,1}\\ &\le n_i^2\|\mathbf{A}_i\|_{P,2}\|\mathbf{B}_i\|_{P,2}\sup_{\kappa,\iota\in[n_i]}|(\mathbf{D}_i)_{\kappa,\iota}|\end{aligned}$$

holds due to the triangle inequality and Hölder's inequality. Because we have $n_i \le n_{\max}$ uniformly over $i \in [N]$, we infer our claim due to

$$\|f_k''(\tilde{r})\| \le \sup_{r\in(0,1)}\Big\|\frac{\partial^2}{\partial r^2}\mathbb{E}_P\left[\varphi\big(\mathbf{S}_i;\Sigma_0+r(\Sigma-\Sigma_0),\eta^0+r(\eta-\eta^0)\big)\right]\Big\| \lesssim \delta_N N^{-\frac{1}{2}}.$$

□

Lemma 3.C.18. *Recall the notation* $\hat{\mathbf{V}}_{i,k} = \mathbf{Z}_i\hat{\Sigma}_k\mathbf{Z}_i^T + \mathbb{1}_{n_i}$. *We have*

$$\frac{1}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T\hat{\mathbf{V}}_{i,k}^{-1}\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k} = T_0 + o_P(1).$$

Proof of Lemma 3.C.18. Let us introduce the score function

$$\xi(\mathbf{S}_i;\Sigma,\eta) := \big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)^T(\mathbf{Z}_i\Sigma\mathbf{Z}_i^T + \mathbb{1}_{n_i})^{-1}\big(\mathbf{X}_i - m_X(\mathbf{W}_i)\big)$$

for $\eta \in \mathcal{T}$ and Σ from Θ. Recall the notation $\mathbf{V}_{i,0} = \mathbf{Z}_i\Sigma_0\mathbf{Z}_i^T + \mathbb{1}_{n_i}$. We have

$$\begin{aligned}&\tfrac{1}{n_{T,k}}\textstyle\sum_{i\in I_k}\left((\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T\hat{\mathbf{V}}_{i,k}^{-1}\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k} - \mathbb{E}_P\left[(\mathbf{R}_{\mathbf{X}_i})^T\mathbf{V}_{i,0}^{-1}\mathbf{R}_{\mathbf{X}_i}\right]\right)\\ =\ &\mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\xi(\mathbf{S};\Sigma_0,\eta^0)]\right]\\ =\ &\mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c}) - \xi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})\right] + \mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c}) - \xi(\mathbf{S};\Sigma_0,\eta^0)\right]\\ &\quad + \mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\Sigma_0,\eta^0) - \mathbb{E}_P[\xi(\mathbf{S};\Sigma_0,\eta^0)]\right].\end{aligned} \tag{3.41}$$

The last summand $\mathbb{E}_{n_{T,k}}[\xi(\mathbf{S};\Sigma_0,\eta^0) - \mathbb{E}_P[\xi(\mathbf{S};\Sigma_0,\eta^0)]]$ in (3.41) is of size $o_P(1)$

due to Markov's inequality and Assumptions 3.B.2.2, 3.B.3.2, and 3.B.4.1. The second summand $\mathbb{E}_{n_{T,k}}[\xi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})-\xi(\mathbf{S};\Sigma_0,\eta^0)]$ in (3.41) is of size $o_P(1)$ due to similar arguments as presented in Lemma 3.C.12. This lemma is stated for a slightly different score function that involves β_0, but the proof of this lemma does not depend on β_0. It can be shown that the same arguments are also valid for the score ξ. The first summand $\mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\hat{\Sigma}_k,\hat{\eta}^{I_k^c})-\xi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})\right]$ in (3.41) is of order $o_P(1)$. To prove this last claim, recall that $\|\hat{\theta}_k-\theta_0\|\leq\delta_N$ holds with P-probability $1-o(1)$ due to Lemma 3.C.7 and because we have $\delta_N^2\leq\delta_N$ for N large enough. Consider Σ from Θ with $\|\Sigma-\Sigma_0\|\leq\delta_N$, and recall the notation $\mathbf{V}_i=\mathbf{Z}_i\Sigma\mathbf{Z}_i^T+\mathbb{1}_{n_i}$. We have

$$
\begin{aligned}
&\mathbb{E}_{n_{T,k}}\left[\xi(\mathbf{S};\Sigma,\hat{\eta}^{I_k^c})-\xi(\mathbf{S};\Sigma_0,\hat{\eta}^{I_k^c})\right]\\
=\;&\tfrac{1}{n_{T,k}}\textstyle\sum_{i\in I_k}\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)^T(\mathbf{V}_i^{-1}-\mathbf{V}_{i,0}^{-1})\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)\\
&+\tfrac{1}{n_{T,k}}\textstyle\sum_{i\in I_k}\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)^T(\mathbf{V}_i^{-1}-\mathbf{V}_{i,0}^{-1})\big(m_X^0(\mathbf{W}_i)-\hat{m}_X^{I_k^c}(\mathbf{W}_i)\big)\\
&+\tfrac{1}{n_{T,k}}\textstyle\sum_{i\in I_k}\big(m_X^0(\mathbf{W}_i)-\hat{m}_X^{I_k^c}(\mathbf{W}_i)\big)^T(\mathbf{V}_i^{-1}-\mathbf{V}_{i,0}^{-1})\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)\\
&+\tfrac{1}{n_{T,k}}\textstyle\sum_{i\in I_k}\big(m_X^0(\mathbf{W}_i)-\hat{m}_X^{I_k^c}(\mathbf{W}_i)\big)^T(\mathbf{V}_i^{-1}-\mathbf{V}_{i,0}^{-1})\big(m_X^0(\mathbf{W}_i)-\hat{m}_X^{I_k^c}(\mathbf{W}_i)\big).
\end{aligned}
\tag{3.42}
$$

The first summand in the decomposition (3.42) is of order $o_P(1)$ because we have for all $i\in[N]$ that

$$
\begin{aligned}
&\mathbb{E}_P\left[\left\|\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)^T(\mathbf{V}_i^{-1}-\mathbf{V}_{i,0}^{-1})\big(\mathbf{X}_i-m_X^0(\mathbf{W}_i)\big)\big)\right\|\right]\\
\leq\;&\sup_{i\in[N]}\|\mathbf{V}_i^{-1}-\mathbf{V}_{0,i}^{-1}\|\|\mathbf{X}_i-m_X^0(\mathbf{W}_i)\|_{P,2}^2\\
\lesssim\;&\delta_N\|\mathbf{X}_i-m_X^0(\mathbf{W}_i)\|_{P,2}^2
\end{aligned}
$$

holds due to the Cauchy-Schwarz inequality, Hölder's inequality, and Assumption 3.B.3.2. We have $\|\mathbf{X}_i-m_X^0(\mathbf{W}_i)\|_{P,2}\leq\|\mathbf{X}_i\|_{P,2}<\infty$ due to Lemma 3.C.2 and Assumption 3.B.2.2. The other summands in (3.42) are of smaller order than the first summand in (3.42) due to Assumption 3.B.4.1 and similar computations. Therefore, we have

$$
\frac{1}{n_{T,k}}\sum_{i\in I_k}(\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k})^T\hat{\mathbf{V}}_{i,k}^{-1}\widehat{\mathbf{R}}_{\mathbf{X}_i}^{I_k}=\frac{1}{n_{T,k}}\sum_{i\in I_k}\mathbb{E}_P\left[(\mathbf{R}_{\mathbf{X}_i})^T\mathbf{V}_i^{-1}\mathbf{R}_{\mathbf{X}_i}\right]+o_P(1)=T_0+o_P(1)
$$

due to Assumption 3.B.2.8. □

3.D | Stochastic Random Effects Matrices

We considered fixed random effects matrices $\mathbf{Z}_i$ in our model (3.2). However, it is also possible to consider stochastic random effects matrices $\mathbf{Z}_i$ and to include the nonparametric variables $\mathbf{W}_i$ into the random effects matrices. In this case, we consider the composite random effects matrices $\widetilde{\mathbf{Z}}_i=\zeta(\mathbf{Z}_i,\mathbf{W}_i)$ for some

known function ζ instead of $\mathbf{Z}_i$ in the partially linear mixed-effects model (3.2). That is, we replace the model (3.2) by the model

$$\mathbf{Y}_i = \mathbf{X}_i\beta_0 + g(\mathbf{W}_i) + \widetilde{\mathbf{Z}}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i \in [N] \tag{3.43}$$

with $\widetilde{\mathbf{Z}}_i = \zeta(\mathbf{Z}_i, \mathbf{W}_i)$ and $\mathbf{Z}_i$ random. We require groupwise independence $\mathbf{Z}_i \perp\!\!\!\perp \mathbf{Z}_j$ for $i \neq j$ of the random effects matrices.

If $\mathbf{Z}_i$ is random, one needs to also condition on it in (3.4), and we need to assume that the density $p(\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i)$ does not depend on θ. Furthermore, $\mathbf{Z}_i$ needs to be such that the Neyman orthogonality properties (3.32) and (3.40) and Equation (3.29) still hold. For instance, these equations remain valid if Assumption 3.2.1.4 is replaced by $(\mathbf{b}_i, \boldsymbol{\varepsilon}_i) \perp\!\!\!\perp (\mathbf{W}_i, \mathbf{X}_i)|\mathbf{Z}_i$ and $\mathbb{E}_P[\mathbf{b}_i|\mathbf{Z}_i] = \mathbf{0}$ and $\mathbb{E}_P[\boldsymbol{\varepsilon}_i|\mathbf{Z}_i] = \mathbf{0}$ for all $i \in [N]$.

Furthermore, the composite random effects matrices $\widetilde{\mathbf{Z}}_i$ need to satisfy additional regularity conditions. The Assumptions 3.B.2.3 and 3.B.3.2 need to be adapted as follows. The first option is to adapt Assumption 3.B.2.3 to: there exists a finite real constant C_2' such that $\|\widetilde{\mathbf{Z}}_i\|_{P,\infty} \leq C_2'$ holds for all $i \in [N]$, where $\|\cdot\|_{P,\infty}$ denotes the $L^\infty(P)$-norm. Then, Assumption 3.B.3.2 needs to be adapted to: there exists a finite real constant C_3' such that we have $\|(\widetilde{\mathbf{Z}}_i\Sigma\widetilde{\mathbf{Z}}_i^T + \mathbb{1}_{n_i})^{-1}\|_{P,\infty} \leq C_3'$. for all $i \in [N]$ and all Σ belonging to Θ.

The Assumptions 3.B.4.1 and 3.B.4.2 formulate the product relationship of the machine learning estimators' convergence rates in terms of the $L^2(P)$-norm. The second option is to consider $L^t(P)$-norms with $t \geq 4 > 2$ in these assumptions instead. Then, it is possible to constrain the $L^p(P)$-norms of $\widetilde{\mathbf{Z}}_i$ and $(\widetilde{\mathbf{Z}}_i\Sigma\widetilde{\mathbf{Z}}_i^T + \mathbb{1}_{n_i})^{-1}$ in Assumptions 3.B.2.3 and 3.B.3.2 instead of their $L^\infty(P)$-norm. However, the order p, which is specified in Assumption 3.B.2, needs to be increased to $p \geq 2^9$ to allow us to bound the terms in the respective proofs by Hölder's inequality.

4 | Regularizing Double Machine Learning in Partially Linear Endogenous Models

Joint work with

Peter Bühlmann

This chapter is based on the manuscript

C. Emmenegger and P. Bühlmann. Regularizing double machine learning in partially linear endogenous models. Electronic Journal of Statistics, 15(2):6461–6543, 2021

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`.

4.1 | 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 (Okui et al., 2012). Another situation where an explanatory variable is correlated with the error term occurs if the explanatory variable is determined simultaneously with the response (Wooldridge, 2013). In such situations, employing estimation methods that do not account for endogeneity can lead to biased estimators (Fuller, 1987).

Let us consider the PLM

$$Y = X^T\beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y. \tag{4.1}$$

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 et al. (2000). Semiparametric methods are summarized in Ruppert et al. (2003) and Härdle et al. (2004), for instance.

Chernozhukov et al. (2018) introduce double machine learning (DML) to estimate the linear coefficient β_0 in a model similar to (4.1). 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 (Pearl, 2009). 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

$$\beta_0 = \frac{\mathbb{E}\left[(A - \mathbb{E}[A|W])(Y - \mathbb{E}[Y|W])\right]}{\mathbb{E}\left[(A - \mathbb{E}[A|W])(X - \mathbb{E}[X|W])\right]} \tag{4.2}$$

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 (Bickel, 1982; Chernozhukov

et al., 2018).

Equation (4.2) has a two-stage least squares (TSLS) interpretation (Theil, 1953a,b; Basmann, 1957; Bowden and Turkington, 1985; Angrist et al., 1996; Anderson, 2005). 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 (Wagner, 1958; Nagar, 1960; Summers, 1965; Cragg, 1967; Lloyd, 1975). The issue with large or nonexisting variance of TSLS (the order of existing moments of TSLS depends on the degree of overidentification (Mariano, 1972, 1982, 2003)) is also coupled with the strength of the instrument (Bound et al., 1995; Staiger and Stock, 1997; Stock et al., 2002; Crown et al., 2011; Andrews et al., 2019). Reducing the variability is sometimes possible by using k-class estimators (Theil, 1961; Hill et al., 2011; Rothenhäusler et al., 2021; Jakobsen and Peters, 2020).

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 et al., 1986) recommend to not use fixed k-class estimators. Three particularly well-established k-class estimators are the limited information maximum likelihood (LIML) estimator (Anderson and Rubin, 1949; Amemiya, 1985) and the Fuller(1) and Fuller(4) estimators (Fuller, 1977). They have been developed for entirely linear models to overcome some deficiencies of TSLS. If many instruments are present, LIML experiences some optimality properties (Anderson et al., 2010). Furthermore, the normal approximation for the finite sample estimator may be suboptimal for TSLS but useful for LIML (Anderson and Sawa, 1979; Anderson et al., 1982; Anderson, 1983). However, LIML has no moments Mariano (1982); Phillips (1984, 1985); Hillier and Skeels (1993). The Fuller estimators overcome this problem. Having no moments can lead to poor squared error performance, especially in weak instrument situations (Hahn et al., 2004). On the other hand, the Fuller(1) estimator is approximately unbiased and Fuller(4) has particularly low mean squared error (MSE) (Fuller, 1977). Takeuchi and Morimune (1985) 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 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.

4.1.1 | Our Contribution

Our contribution is twofold. First, we build on the work of Chernozhukov et al. (2018) to estimate β_0 in the endogenous PLM (4.1) 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. (2018), 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 (Amemiya, 1974; Berndt et al., 1974; Hansen, 1985) and more robust estimators (Pearl, 2004). Considering SEMs and an identifiability condition allows us to apply DML to more general situations than in Chernozhukov et al. (2018).

Second, we propose a DML method that employs regularization and selection. This method is called regsDML, and we develop it in Section 4.4. 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 (Theil, 1961) and anchor regression (Rothenhäusler et al., 2021; Bühlmann, 2020). 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 (2020) 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 (2020), 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

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

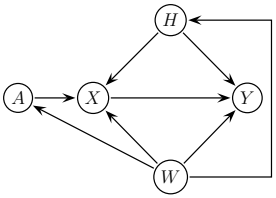


Figure 4.1.1: An SEM and its associated causal graph.

enters linearly in (4.1) 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 (Bang and Robins, 2005). The detailed assumptions on how the variables A, X, W, H, and Y interact are given in Section 4.2: 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 4.1.1 gives the SEM we generate data from and its associated causal graph (Lauritzen, 1996; Pearl, 1998, 2009, 2010; Peters et al., 2017; Maathuis et al., 2019). By convention, we omit error variables in a causal graph if they are mutually independent (Pearl, 2009). The variable A is similar to a conditional instrument given W.

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 4.1.2. 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 method is a version of regsDML with regularization only but no selection. If the blue curve is not visible in Figure 4.1.2, 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

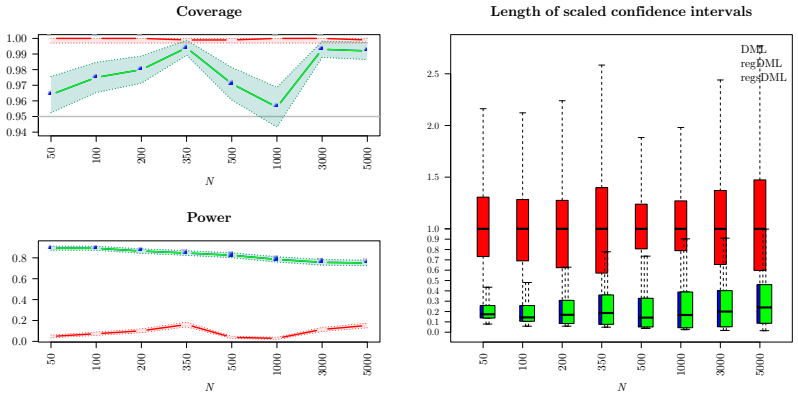


Figure 4.1.2: The results come from $M = 1000$ simulation runs each from the SEM in Figure 4.1.1 for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided confidence 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.

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 4.1.2 are presented in Figure 4.D.1 in Section 4.D in the appendix. Further numerical results are given in Section 4.5.

4.1.2 | Additional Literature

PLMs have received considerable interest. Härdle et al. (2000) 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 (2003) consider semiparametric estimation with a sieve estimator. Ma and Carroll (2006) introduce a parametric model for the latent variable. Yao (2012) considers a heteroskedastic error term and a partialling-out scheme (Robinson, 1988; Speckman, 1988).

Florens et al. (2012) propose to solve an ill-posed integral equation. Su and Zhang (2016) 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 (2011) 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 (1994), van der Laan and Robins (2003), and Chernozhukov et al. (2018). DML methods have been applied in various situations. Chen et al. (2021) consider instrumental variables quantile regression. Liu et al. (2021) apply DML in logistic partially linear models. Colangelo and Lee (2020) employ doubly debiased machine learning methods to a fully nonparametric equation of the response with a continuous treatment. Knaus (2020) presents an overview of DML methods in unconfounded models. Farbmacher et al. (2020) decompose the causal effect of a binary treatment by a mediation analysis and estimate it by DML. Lewis and Syrgkanis (2020) extend DML to estimate dynamic effects of treatments. Chiang et al. (2021) apply DML under multiway clustered sampling environments. Cui and Tchetgen Tchetgen (2020) 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 (1997). Alternatively, Chen et al. (2016) partial out the nonparametric component using a kernel method and employ the generalized method of moments principle (Hansen, 1982). 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 (Chernozhukov et al., 2018).

Double robustness and orthogonality arguments have also been considered in the following works. Okui et al. (2012) 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 et al. (2019) 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 et al. (2019).

The literature presented in this subsection is related to but rather distinct from our work with the only exception of Chernozhukov et al. (2018). The difference to this latter contribution is highlighted in Section 4.2 and Section 4.A in the appendix.

Outline of the Paper. Sections 4.2 and 4.3 describe the DML estimator. The former section introduces an identifiability condition, and the latter investigates asymptotic properties. Section 4.4 introduces the regularized regularization-selection estimator regDML and its regularization-only version regDML and investigates their asymptotic properties. Section 4.5 presents numerical experiments and an empirical real data example. Section 4.6 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 $\xrightarrow{d}$ 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.

4.2 | An Identifiability Condition and the DML Estimator

Before we introduce regsDML in Section 4.4, 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. (2018), 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

$$Y \leftarrow X^T\beta_0 + g_Y(W) + h_Y(H) + \varepsilon_Y \tag{4.3}$$

as in (4.1), 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 (4.3). The model is required to satisfy an identifiability condition as in (4.5) 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 (Anderson, 1983).

4.2.1 | Identifiability Condition

An identifiability condition is required to identify β_0 in (4.3). We define the residual terms

$$R_A := A - \mathbb{E}[A|W], \quad R_X := X - \mathbb{E}[X|W], \quad \text{and} \quad R_Y := Y - \mathbb{E}[Y|W] \tag{4.4}$$

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

$$\mathbb{E}\left[R_A(R_Y - R_X^T\beta_0)\right] = \mathbf{0} \tag{4.5}$$

to identify β_0 in (4.3). 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 (4.5) is strictly weaker than the conditional moment conditions introduced in Chernozhukov et al. (2018); see Section 4.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 4.2.1. *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* (4.5)*. We then have*

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

Theorem 4.2.1 precludes underidentification. The full rank condition of the matrix $\mathbb{E}_P[R_X R_A^T]$ expresses that the correlation between X and A is strong enough after regressing out W. This is a typical TSLS assumption (Theil, 1953a,b; Basmann, 1957; Bowden and Turkington, 1985; Angrist et al., 1996; Anderson, 2005). The rank assumptions in Theorem 4.2.1 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 4.2.1. However, this procedure leads to a suboptimal convergence rate of the resulting estimator; see Section 4.3.1.

The identifiability condition (4.5) is central to Theorem 4.2.1. Section 4.G in the appendix presents examples illustrating SEMs where the identifiability condition holds and where it fails to hold.

4.2.2 | 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 4.4. 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

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

By Theorem 4.2.1, the population parameter β_0 solves the TSLS moment equation

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

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

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

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

$$\beta_0 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T \beta)\right)^2\right]. \tag{4.6}$$

We employ the representation of β_0 in (4.6) to formulate our regularization schemes in Section 4.4.

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 (4.3). 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 $I_1, \ldots, I_K$. 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

$$\hat{\beta} := \left(\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\right)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\right)^{-1} \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\right)^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \tag{4.7}$$

of β_0, where

$$\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} := \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\left(\left(\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\right)^T \widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\right)^{-1}\left(\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}\right)^T \tag{4.8}$$

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 (4.7), 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. (2018) call these two schemes DML2 and DML1, respectively, where DML2 is as in (4.7). The DML1 version of the coefficient estimator is given in the appendix in Section 4.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. (2018). This procedure is described in Algorithm 3 in Section 4.4.2 below.

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

Theorem 4.3.1. *Consider model* (4.3). *Suppose that Assumption 4.I.5 in the appendix in Section 4.I holds and consider* $\overline{\psi}$ *given in Definition 4.I.1 in the appendix in Section 4.I. Then* $\hat{\beta}$ *as in* (4.7) *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* $\sigma^2 = J_0 \tilde{J}_0 J_0^T$ *for the matrices* $\tilde{J}_0$ *and* J_0 *given in Definition 4.I.1 in the appendix.*

A similar result to Theorem 4.3.1 is presented by Chernozhukov et al. (2018). However, their result requires univariate A and X, and it imposes conditional moment restrictions instead of the identifiability condition (4.5); see also Section 4.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. (2018).

Theorem 4.3.1 also holds for the DML1 version of $\hat{\beta}$ defined in the appendix in Section 4.B.1. Assumption 4.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

$$\begin{aligned} &\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2}^2 \ll N^{-\frac{1}{2}}, \\ &\|m_A^0(W) - \hat{m}_A^{I_k^c}(W)\|_{P,2} \\ &\quad \cdot \big(\|m_Y^0(W) - \hat{m}_Y^{I_k^c}(W)\|_{P,2} + \|m_X^0(W) - \hat{m}_X^{I_k^c}(W)\|_{P,2}\big) \ll N^{-\frac{1}{2}} \end{aligned} \tag{4.9}$$

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 (Candes and Tao, 2007; Bickel et al., 2009; Bühlmann and van de Geer, 2011; Belloni and Chernozhukov, 2013), forward selection in sparse linear models (Kozbur, 2020), high-dimensional additive models (Meier et al., 2009; Koltchinskii and Yuan, 2010; Yuan and Zhou, 2016), or regression trees and random forests (Wager and Walther, 2016; Athey et al., 2019). Please see Chernozhukov et al. (2018) for additional references. In particular, the rate condition (4.9) 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 4.I.1 and in Theorem 4.I.21 in the appendix in Section 4.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 4.3.1 holds uniformly over laws P. This uniformity guarantees some robustness of the asymptotic statement (Chernozhukov et al., 2018). 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

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

where the score function ψ is given by

$$\psi(S;\beta,\eta) := \big(A - m_A(W)\big)\Big(Y - m_Y(W) - \big(X - m_X(W)\big)^T\beta\Big) \quad (4.10)$$

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 (4.5). 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 (4.9). 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 (4.10) 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 4.3.2. *(Chernozhukov et al., 2018, Definition 2.1). A score $\psi = \psi(S; \beta, \eta)$ is Neyman orthogonal at (β_0, η^0) if the pathwise derivative map*

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

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

Definition 4.3.2 does not entirely coincide with Chernozhukov et al. (2018, Definition 2.1) because the latter also includes an identifiability condition. We directly assume the identifiability condition (4.5).

The subsequent proposition states that the score function ψ in (4.10) is indeed Neyman orthogonal.

Proposition 4.3.3. *The score ψ given in Equation* (4.10) *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 the coefficients β_0 and η^0. In addition to being Neyman orthogonal, ψ is linear in β in the sense that we have

$$\psi(S; \beta, \eta) = \psi^b(S; \eta) - \psi^a(S; \eta)\beta \tag{4.11}$$

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 4.3.1.

4.3.1 | 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

$$\frac{1}{\sqrt{n}} \sum_{i \in I} A_i \big(\widehat{R}^I_{Y,i} - (\widehat{R}^I_{X,i})^T \beta_0\big) \tag{4.12}$$

can diverge as $N \to \infty$ because $\hat{m}^{I^c}_X$ and $\hat{m}^{I^c}_Y$ can 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. (2018). Even if sample splitting is employed, the term (4.12) is

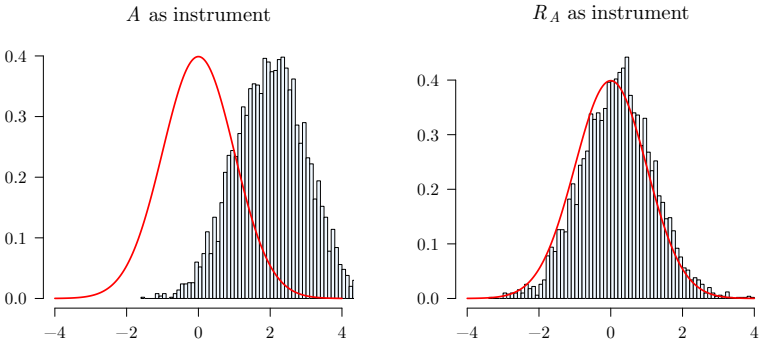


Figure 4.3.1: Histograms of $\frac{\hat{\beta}-\beta_0}{\widehat{\text{Var}}(\hat{\beta})}$, where $\widehat{\text{Var}}(\hat{\beta})$ denotes the empirically observed variance of $\hat{\beta}$ with respect to the simulation runs, using A as an instrument in the left plot and using R_A as an instrument 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 4.C with $K = 2$. The conditional expectations are estimated with random forests consisting of 500 trees that have a minimal node size of 5.

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 4.3.1. The SEM used to generate the data is similar to the nonconfounded model used in Chernozhukov et al. (2018, Figure 1). The centered and rescaled term $\frac{\hat{\beta}-\beta_0}{\widehat{\text{Var}}(\hat{\beta})}$ using A as an instrument is biased whereas it is not if the instrument R_A is used. Here, $\widehat{\text{Var}}(\hat{\beta})$ denotes the empirically observed variance of $\hat{\beta}$ with respect to the performed simulation runs.

4.4 | 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 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 estimators 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 (Theil, 1961) or anchor regression (Rothenhäusler et al., 2021; Bühlmann, 2020). 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 (4.6). For a regularization parameter $\gamma \geq 0$, we consider the regularized objective function and corresponding population coefficient

$$b^\gamma := \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\left((\mathrm{Id} - P_{R_A})(R_Y - R_X^T\beta)\right)^2\right] + \gamma\, \mathbb{E}\left[\left(P_{R_A}(R_Y - R_X^T\beta)\right)^2\right]. \tag{4.13}$$

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 (2020) 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 (2020).

4.4.1 | Estimation and Asymptotic Normality

In this section, we describe how to estimate b^γ in (4.13) 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 4.3 that adjust the data with respect to the

nonparametric variables. The estimator of b^γ is given by

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

where $\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}$ is as in (4.8). This estimator can be expressed in closed form by

$$\hat{b}^\gamma = \left(\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}\right)^T\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}\right)^{-1}\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}\right)^T\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k}, \tag{4.14}$$

where

$$\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k} := \left(\mathbb{1}+(\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\right)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \quad\text{and}\quad \widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k} := \left(\mathbb{1}+(\sqrt{\gamma}-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\right)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}. \tag{4.15}$$

The computation of $\hat{b}^\gamma$ is similar to an OLS scheme where $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k}$ is regressed on $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}$. 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 $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k}$ on $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}$ 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. (2018) as already mentioned in Section 4.3. The DML1 version is presented in the appendix in Section 4.B.2. As mentioned in Section 4.3, 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 4.4.1. *Let $\gamma \geq 0$. Suppose that Assumption 4.I.5 in the appendix in Section 4.I (same as in Theorem 4.3.1) except 4.I.5.1 holds, and consider the quantities $\sigma^2(\gamma)$ and $\overline{\psi}$ introduced in Definition 4.J.1 in the appendix in Section 4.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 4.4.1 also holds for the DML1 version of $\hat{b}^\gamma$ defined in the appendix in Section 4.B.2. The influence function is denoted by $\overline{\psi}$ in both Theorems 4.3.1 and 4.4.1 but is defined differently. Assumption 4.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

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

for $k \in [K]$. The main difference to Theorem 4.3.1 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 4.J.3 in the appendix in Section 4.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 4.4.1 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 4.3.

For fixed $\gamma > 1$, Theorem 4.4.1 furthermore implies that the k-class estimator corresponding to $\hat{b}^\gamma$ converges at the parametric rate and follows a Gaussian distribution asymptotically.

4.4.2 | 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. This scheme first optimizes the estimated asymptotic MSE of $\hat{b}^\gamma$. The estimated regularization for the parameter γ leads to an estimate of β_0 that asymptotically has the same MSE behavior as the TSLS-type estimator $\hat{\beta}$ in (4.7) but may exhibit substantially better finite sample properties.

We consider the estimated regularization parameter

$$\hat{\gamma} := \arg\min_{\gamma \geq 0} \frac{1}{N} \hat{\sigma}^2(\gamma) + |\hat{b}^\gamma - \hat{\beta}|^2. \tag{4.16}$$

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 4.J.3 in the appendix in Section 4.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 and analyze the asymptotic properties of regDML and regsDML.

Let us consider a deterministic sequence $\{\gamma_N\}_{N\geq 1}$ of regularization parameters. By Proposition 4.4.2 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 4.4.2. *Suppose that 4.I.5.1, 4.I.5.3, and 4.I.5.4 of Assumption 4.I.5 in the appendix in Section 4.I hold (subset of the assumptions in Theorem 4.3.1). Assume $\{\gamma_N\}_{N\geq 1}$ is 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 4.4.3 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 (4.16), and hence selecting the TSLS-type estimator $\hat{\beta}$, would lead to a smaller estimated asymptotic MSE.

Theorem 4.4.3. *Let $\gamma_N = o(\sqrt{N})$, and suppose that Assumption 4.I.5 in the appendix in Section 4.I holds (same as in Theorem 4.3.1). 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 4.4.2 is formulated for deterministic regularization parameters, but the deterministic statements can be replaced by probabilistic ones. Proposition 4.4.2 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 4.4.4 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 4.4.4).

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

Theorem 4.4.4. *Suppose that Assumption 4.I.5 in the appendix in Section 4.I holds (same as in Theorem 4.3.1). 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 4.J.3 in the appendix, which consistently estimates $\sigma(\cdot)$ from 4.4.1.

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

The proof of Theorem 4.4.4 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 4.4.4 remains valid if the regularization parameter comes from k-class estimaton and is of the required stochastic order. The same stochastic order is also required to show that k-class estimators are asymptotically Gaussian (Nagar, 1959; Mariano, 2003).

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. (2018). The assembled parameter estimate is given by the median of the individual parameter estimates; see Steps 9 and 10 of Algorithm 3. 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 3.

Algorithm 3: regsDML in a PLM with confounding variables.

Input : N iid realizations from the SEM (4.3), a natural number $\mathcal{S}$, a regularization parameter grid $\{\gamma_i\}_{i\in[M]}$ for some natural number M, a non-negative diverging sequence $\{a_n\}_{n\geq 1}$.

Output : An estimator of β_0 in (4.3) 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 $i \in [M]$.
  Choose $\hat{\gamma}_s = \arg\min_{\gamma\in\{\gamma_i\}_{i\in[M]}} \left(\frac{1}{N}\hat{\sigma}_s^2(\gamma) + |\hat{b}_s^{\gamma} - \hat{\beta}_s|^2\right)$ 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 $\hat{\beta}^{\text{med}} = \text{median}_{s\in[\mathcal{S}]}(\hat{\beta}_s)$.
Compute $\hat{b}_{\text{reg}}^{\text{med}} = \text{median}_{s\in[\mathcal{S}]}(\hat{b}_s^{\hat{\gamma}_s'})$.
Compute $\hat{\sigma}^{2,\text{med}} = \text{median}_{s\in[\mathcal{S}]}\big(\hat{\sigma}_s^2 + (\hat{\beta}_s - \hat{\beta}^{\text{med}})^2\big)$.
Compute $\hat{\sigma}_{\text{reg}}^{2,\text{med}} = \text{median}_{s\in[\mathcal{S}]}\big(\hat{\sigma}_s^2(\hat{\gamma}_s') + (\hat{b}_s^{\hat{\gamma}_s'} - \hat{b}_{\text{reg}}^{\text{med}})^2\big)$.
if $\hat{\sigma}_{\text{reg}}^{2,\text{med}} < \hat{\sigma}^{2,\text{med}}$ **then**
  Take the parameter estimate $\hat{b}_{\text{reg}}^{\text{med}}$ together with its associated estimated asymptotic variance $\frac{1}{N}\hat{\sigma}_{\text{reg}}^{2,\text{med}}$.
else
  Take the parameter estimate $\hat{\beta}^{\text{med}}$ together with its associated estimated asymptotic variance $\frac{1}{N}\hat{\sigma}^{2,\text{med}}$.
end

4.5 | 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` (Emmenegger, 2021). We employ the DML2 method and $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. 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 3.

The first example in Section 4.5.1 considers an overidentified model in which the dimension of A is larger than the dimension of X. The conditional expectations acting as nuisance parameters are estimated with random forests. The

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

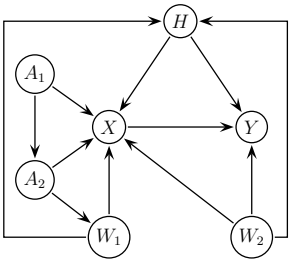


Figure 4.5.1: An SEM and its associated causal graph.

second example in Section 4.5.2 considers justidentified real-world data. The conditional expectations are also estimated with random forests.

An example where the conditional expectations are estimated with splines is given in Section 4.1.1. Additional empirical results are provided in the appendix in Sections 4.D, 4.E, and 4.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 (Bühlmann and van de Geer, 2018) to possibly enhance further insights into the finite sample behavior. Section 4.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 γ.

4.5.1 | Simulation Example with Random Forests

We generate data from the SEM in Figure 4.5.1. This SEM satisfies the identifiability condition (4.5) 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 4.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.

We simulate $M = 1000$ datasets each from the SEM in Figure 4.5.1 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 3 and estimate the conditional expectations with random forests consisting of 500 trees that have a minimal node size of 5.

Figure 4.5.2 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 4.5.2, 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 curves are not visible, they coincide 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 4.5.2 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 eft plot in Figure 4.5.2 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 4.5.2 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.

Simulation results with $\beta_0 = 0$ in the SEM in Figure 4.5.1 are presented in Figure 4.D.2 in the appendix in Section 4.D.

4.5.2 | 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 et al. (2001) and Chernozhukov et al. (2018). 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 et al. (2001) and Chernozhukov et al. (2018). The data is available in the R-package hdm (Chernozhukov et al., 2016) 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

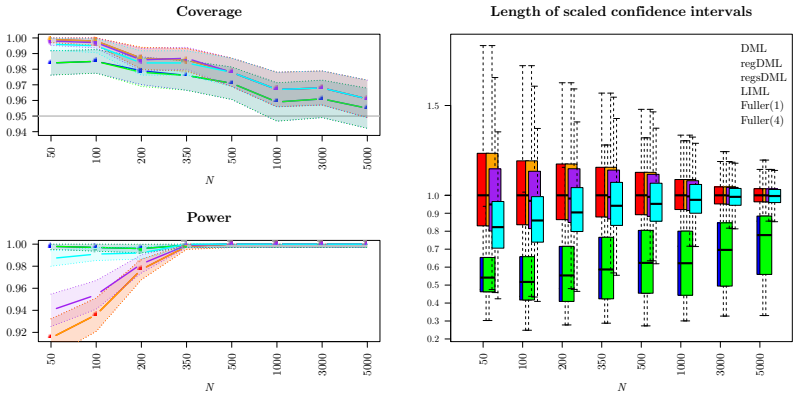


Figure 4.5.2: The results come from $M = 1000$ simulation runs each from the SEM in Figure 4.5.1 for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with random forests. The figure displays the coverage of two-sided confidence intervals 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.

	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	$[0.239, 1.136]$

Table 4.5.1: Coefficient estimate, its standard error, and a confidence interval with regsDML and DML on the AJR dataset, where $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3, and where the conditional expectations are estimated with random forests consisting of 1000 trees that have a minimal node size of 5.

factors Africa, Asia, North America, and South America. That is, we adjust nonparametrically for the latitude and geographic information.

We choose $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3 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 4.5.1. 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. (2018). 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. (2017). 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.

4.6 | 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 (Okui et al., 2012). In such situations, employing estimation methods that do not account for endogeneity lead to biased estimators (Fuller, 1987).

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. (2018). However, our setting is more general than the one considered in Chernozhukov et al. (2018) 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 (Theil, 1961) and anchor regression (Rothenhäusler et al., 2021; Bühlmann, 2020; Jakobsen and Peters, 2020) 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 4.F in the appendix presents examples with strong and reduced confounding and demonstrates the coverage behavior of DML and regsDML. Section 4.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` (Emmenegger, 2021).

Acknowledgements

We thank Matthias Löffler for constructive comments.

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).

Appendix

4.A | An Example where the Identifiability Condition (4.5) holds, but Conditional Moment Requirements do not

This section presents an SEM where our identifiability condition (4.5) holds, but where the conditional moment requirements of Chernozhukov et al. (2018) do not.

We assume the model

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

given in (4.3) and the identifiability condition $\mathbb{E}_P[R_A(R_Y - R_X^T\beta_0)] = \mathbf{0}$ given in (4.5). Chernozhukov et al. (2018) assume the model

$$Y = X^T\beta_0 + g_Y(W) + U, \quad A = g_A(W) + V \tag{4.17}$$

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

$$\mathbb{E}[U|A, W] = 0 \quad \text{and} \quad \mathbb{E}[V|W] = \mathbf{0} \tag{4.18}$$

on the error terms. Their model is implicitly assumed to be justidentified: the dimensions of A and X are implicitly assumed to be equal.

Model (4.17) and the conditional moment restrictions (4.18) imply the identifiability condition (4.5) due to

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

However, the reverse direction does not hold. A counterexample is presented in Figure 4.A.1 where W directly affects H. This SEM satisfies the identifiability condition (4.5) 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.

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

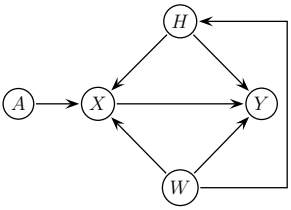


Figure 4.A.1: An SEM and its associated causal graph.

4.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.

4.B.1 | DML1 Estimator of β_0

The DML1 estimator of β_0 is given by

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

where

$$\hat{\beta}^{I_k} := \left((\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k} \right)^{-1} (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} \widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k}, \tag{4.19}$$

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 (4.8). 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 $\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}$.

4.B.2 | DML1 estimator of b^γ

The DML1 estimator of b^γ is given by

$$\hat{b}^{\gamma,\text{DML1}} := \frac{1}{K} \sum_{k=1}^{K} \hat{b}_k^{\gamma}, \tag{4.20}$$

where

$$\hat{b}_k^\gamma := \underset{b\in\mathbb{R}^d}{\arg\min} \left(\left\| (\mathbb{1} - \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}) (\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T b) \right\|_2^2 + \gamma \left\| \Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}} (\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} - (\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k})^T b) \right\|_2^2 \right).$$

This estimator can be expressed in closed form by

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

where we recall the notation

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

as in (4.15). The computation of $\hat{b}_k^\gamma$ is an OLS scheme where $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{Y}}}^{I_k}$ is regressed on $\widehat{\boldsymbol{R}}_{\widetilde{\boldsymbol{X}}}^{I_k}$.

4.C | SEM of Figure 4.3.1

The data from the simulation displayed in Figure 4.3.1 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

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

4.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 4.D.1 and 4.D.2 illustrate the simulation results with $\beta_0 = 0$ of the examples presented in Figure 4.1.2 and 4.5.2 in Sections 4.1.1 and 4.5.1, 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 4.1.2 and 4.5.2, Figure 4.D.1 and 4.D.2 illustrate the type I error.

In Figure 4.D.1, 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 4.D.1 indicates that the lengths 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 4.D.2, the type I errors of both DML and regsDML are similar. The 95% confidence regions of both estimators include the 5% level or are below it. The 95% confidence regions of the levels are represented by dashed lines. These confidence regions of both DML and regsDML contain the 5% level or are below it. The right plot in Figure 4.D.2 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.

4.E | Weak $A \to X$ and Bias-Variance Tradeoff

First, we analyze the behavior of our methods for varying strength of A on 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 4.1.2 and 4.5.2.

Figure 4.E.1 illustrates the results for data from the SEM from Figure 4.1.2. We vary the strength of the direct link $A \to X$ and denote it by α in Figure 4.E.1. Figure 4.E.2 illustrates the results for data from the SEM from Figure 4.5.2. 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 4.E.2. In both Figure 4.E.1 and 4.E.2, the coverage remains high for all considered methods. If α becomes larger, the confidence intervals become shorter, which leads to a coverage that is closer to the nominal 95% level, especially in Figure 4.E.2. 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 4.1.2 and 4.5.2. The results are summarized in Figure 4.E.3 and 4.E.4 that display the estimated MSE, estimated

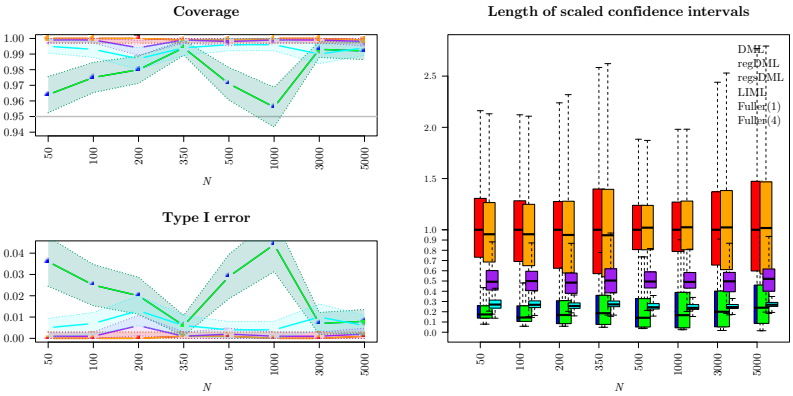


Figure 4.D.1: The results come from $M = 1000$ simulation runs each from the SEM in Figure 4.1.1 with $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage of two-sided 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.

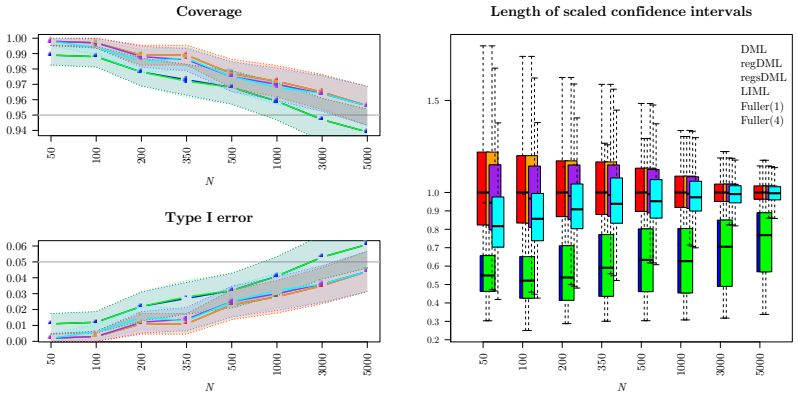


Figure 4.D.2: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.5.1 with $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with random forests. The figure displays the coverage of two-sided 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.

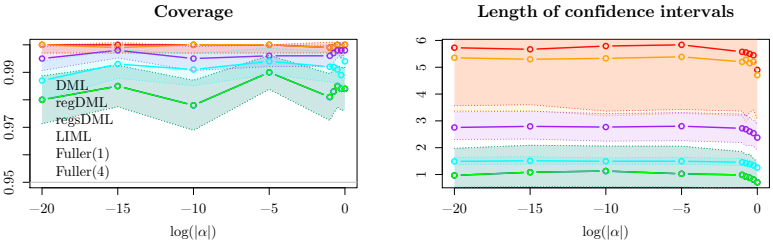


Figure 4.E.1: Same setting as in Figure 4.1.2, but with $N = 200$ only. The strength of the direct link $A \to X$ varies and is denoted by α.

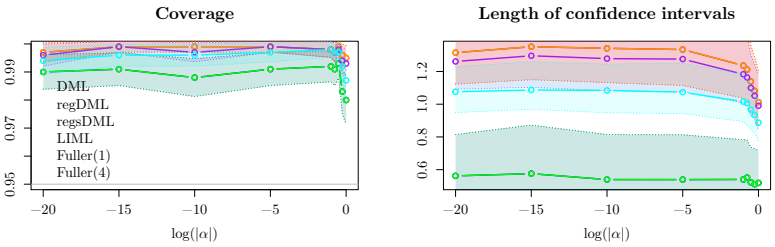


Figure 4.E.2: Same setting as in Figure 4.5.2, but with $N = 200$ only. The strength of the direct link $A_1 \to X$ varies and is denoted by α.

variance, and estimated squared bias as used in Equation (4.16). 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.

4.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 $\lceil N^{\frac{1}{5}} \rceil + 2$ 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.

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 4.F.1 features a strong confounding effect $H \to X$, the second model in Section 4.F.2 features an effect with noise in $W \to H$, and the third model in Section 4.F.3 features an effect with noise in $H \to W$.

4.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 4.F.2.

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 (4.16). The

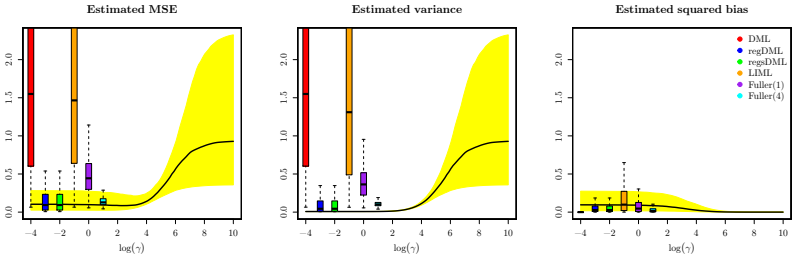


Figure 4.E.3: Estimated MSE, estimated variance, and estimated squared bias as used in Equation (4.16) for the same setting as in Figure 4.1.2, but with $N = 200$ only. The black solid line displays the median of the respective quantity over the considered range of γ-values for $\hat{b}^{\gamma}$. The yellow area marks the observed 25% and 75% quantiles. All methods incorporate an additional variance adjustment from the $\mathcal{S}$ repetitions according to Algorithm 3. 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.

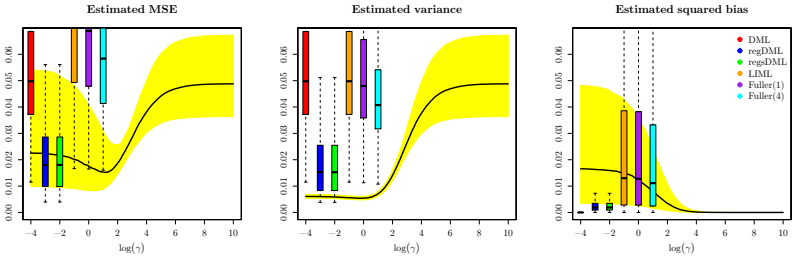


Figure 4.E.4: Estimated MSE, estimated variance, and estimated squared bias as used in Equation (4.16) for the same setting as in Figure 4.5.2, but with $N = 200$ only. The black solid line displays the median of the respective quantity over the considered range of γ-values for $\hat{b}^{\gamma}$. The yellow area marks the observed 25% and 75% quantiles. All methods incorporate an additional variance adjustment from the $\mathcal{S}$ repetitions according to Algorithm 3. 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.

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

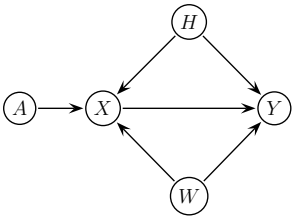


Figure 4.F.1: An SEM and its associated causal graph.

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 4.F.3.

4.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 4.F.5. 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 4.F.6.

4.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 4.F.2. The results again depend on the choice of functions in the model.

Figure 4.F.8 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 4.F.9.

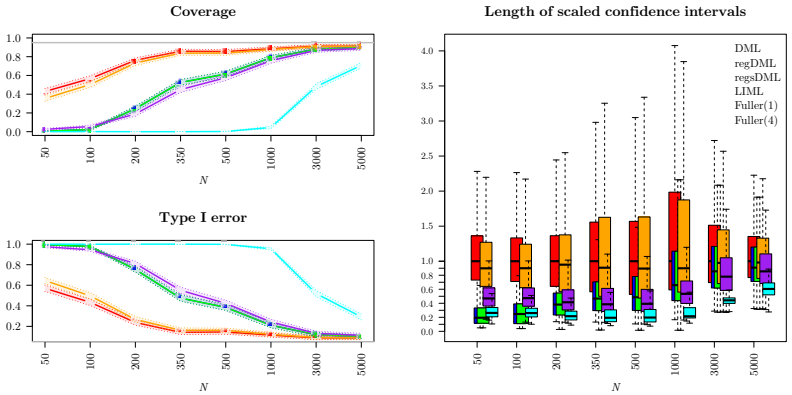


Figure 4.F.2: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.1 with $\chi = 15$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage 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.

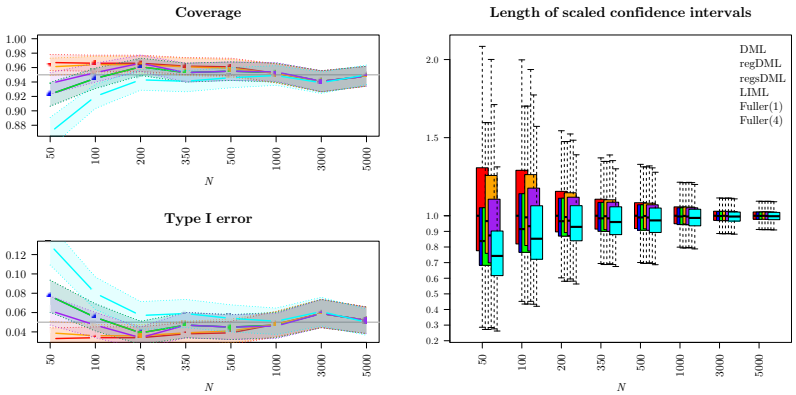


Figure 4.F.3: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.1 with $\chi = 1$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage 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.

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

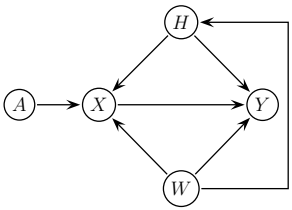


Figure 4.F.4: An SEM and its associated causal graph.

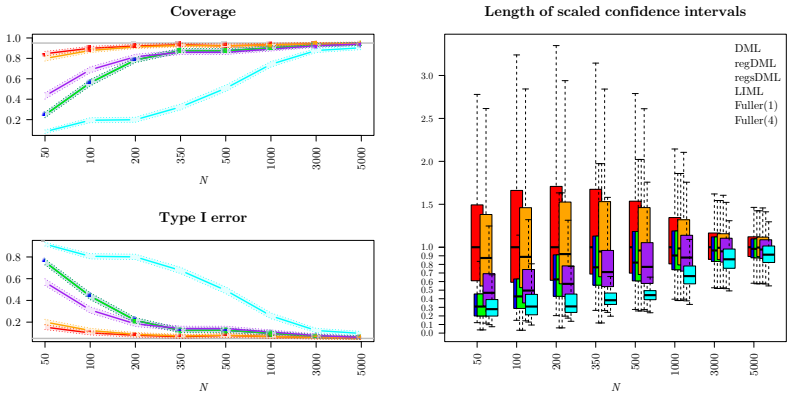


Figure 4.F.5: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.4 with $\kappa = 2$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage 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.

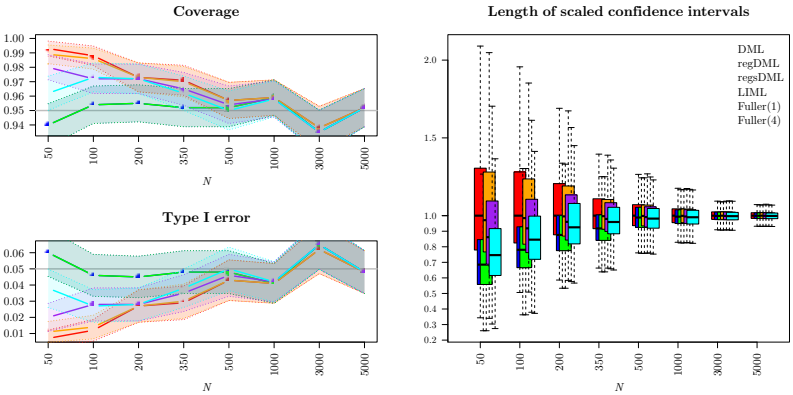


Figure 4.F.6: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.4 with $\kappa = 0.25$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The figure displays the coverage of two-sided confidence intervals for β_0, type I 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.

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

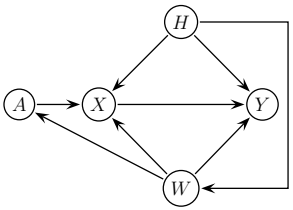


Figure 4.F.7: An SEM and its associated causal graph.

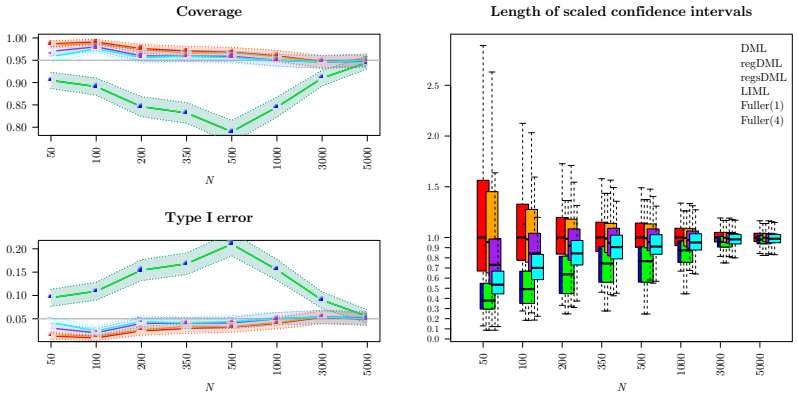


Figure 4.F.8: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.7 with $\kappa = 1$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the coverage 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.

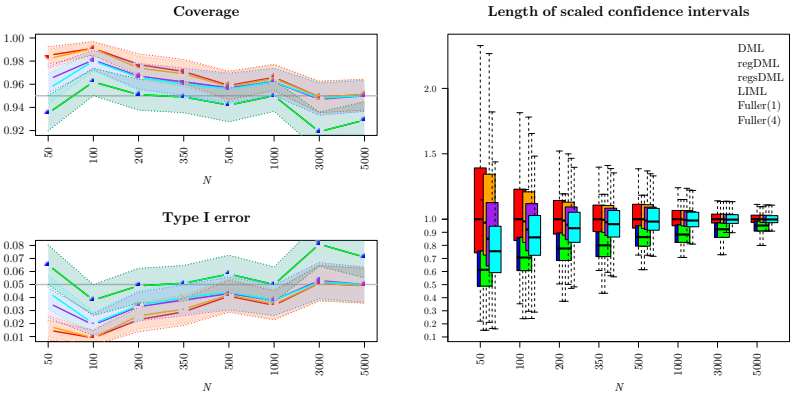


Figure 4.F.9: The results come from $M = 1000$ simulation runs from the SEM in Figure 4.F.7 with $\kappa = 0.25$ and $\beta_0 = 0$ for a range of sample sizes N and with $K = 2$ and $\mathcal{S} = 100$ in Algorithm 3. The nuisance functions are estimated with additive splines. The figure displays the 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.

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

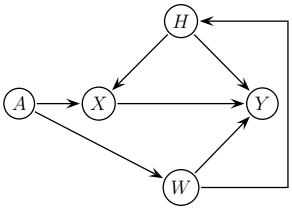


Figure 4.G.1: An SEM satisfying the identifiability condition (4.5) and its associated causal graph as in Example 4.G.1.

4.G | Examples where the identifiability condition (4.5) does and does not hold

The following examples illustrate SEMs where the identifiability condition (4.5) holds and where it fails to hold. We argue using causal graphs; see Lauritzen (1996); Pearl (1998, 2009, 2010); Peters et al. (2017); Maathuis et al. (2019). By convention, we omit error variables in a causal graph if they are assumed to be mutually independent (Pearl, 2009).

Example 4.G.1. *Consider the SEM of the 1-dimensional variables A, W, H, X, and Y and its associated causal graph given in Figure 4.G.1, 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; a proof is given in the appendix in Section 4.H.*

Proof of Example 4.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 4.G.1. In general, this is no requirement; see Example 4.G.2.

Example 4.G.2. *Consider the SEM of the 1-dimensional variables H, W, A, X, and Y and its associated causal graph given in Figure 4.G.2, 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 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; a proof is given in the appendix in Section 4.H.*

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

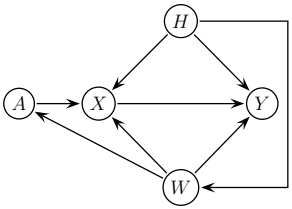


Figure 4.G.2: An SEM satisfying the identifiability condition (4.5) and its associated causal graph as in Example 4.G.2.

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

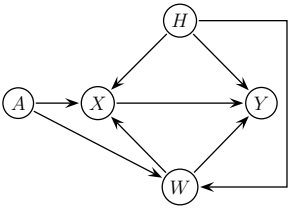


Figure 4.G.3: An SEM not satisfying the identifiability condition (4.5) together with its associated causal graph as in Example 4.G.3

Proof of Example 4.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 4.G.3. Considering the instrument A instead of R_A in Theorem 4.2.1 cannot solve the issue. In such a situation, stronger structural assumptions are required.

Example 4.G.3. *Consider the SEM of the 1-dimensional variables H, A, W, X, and Y and its associated causal graph given in Figure 4.G.3, where β_0 is a fixed unknown parameter. Although A and H are independent, the identifiability condition* (4.5) *does not hold; a proof is given in the appendix in Section 4.H.*

Proof of Example 4.G.3. The two random variables A and H are independent 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 4.2.1 is equivalent to the identifiability condition

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

in Equation (4.5). However, the identifiability condition does not hold in the present situation. We have

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

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 4.2.1. □

4.H | Proofs of Section 4.2

Proof of Theorem 4.2.1. To prove the theorem, we need to verify

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

This statement is equivalent to

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

This last statement holds because $\mathbb{E}[R_A(R_Y - R_X^T\beta_0)]$ equals $\mathbf{0}$ due to the identifiability condition (4.5). □

4.I | Proofs of Section 4.3

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

Proof of Proposition 4.3.3. We have

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

Subsequently, we show that both terms

$$\mathbb{E}_P\left[\big(m_A(W)-m_A^0(W)\big)\Big(Y-m_Y^0(W)-\big(X-m_X^0(W)\big)^T\beta_0\Big)\right] \tag{4.21}$$

and

$$\mathbb{E}_P\left[\big(A-m_A^0(W)\big)\Big(-\big(m_Y(W)-m_Y^0(W)\big)+\big(m_X(W)-m_X^0(W)\big)^T\beta_0\Big)\right] \tag{4.22}$$

are equal to $\mathbf{0}$. We first consider the term (4.21). Recall the notations $m_Y^0(W) = \mathbb{E}_P[Y|W]$ and $m_X^0(W) = \mathbb{E}_P[X|W]$. We have

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

Next, we verify that the term given in (4.22) vanishes. Recall the notation $m_A^0(W) = \mathbb{E}_P[A|W]$. We have

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

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

Because both terms (4.21) and (4.22) vanish, we conclude

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

□

Definition 4.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*

$$
\widetilde{\psi}(S, \beta, \eta) := \big(X - m_X(W)\big)\Big(Y - m_Y(W) - \big(X - m_X(W)\big)^T \beta\Big), \quad (4.23)
$$

and

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

Furthermore, let the matrices

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

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

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

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

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 4.I.3. *We would like to remark that the respective definition of the statistical rate* r_N *given in Chernozhukov et al. (2018) 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 4.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 4.I.16.*

Definition 4.I.4. *Let the nonrandom numbers*

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

If not stated otherwise, we assume the following Assumption 4.I.5 in all the results presented in the appendix.

Assumptions 4.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 sets* $\mathcal{T}$ *such that the following conditions hold:*

4.I.5.1 We have an SEM given by (4.3) *that satisfies the identifiability conditon* (4.5).

4.I.5.2 There exists a finite real constant C_1 *satisfying* $\|A\|_{P,p} + \|X\|_{P,p} + \|Y\|_{P,p} \leq C_1$.

4.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* $c_2 < \infty$.

4.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 4.J.1 in the appendix in Section 4.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.

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

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

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*

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

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 4.I.6. *It is possible to drop some of the assumptions in Assumption 4.I.5 if we are interested in proving the results about DML only. The full assumption is required to prove the results about both DML and regDML.*

We assume Assumption 4.I.5 throughout.

Lemma 4.I.7. *Let* $u \geq 1$*. Consider a* t*-dimensional random variable* Z*. Denote the joint law of* Z *and* W *by* P*. Then we have*

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

Proof of Lemma 4.I.7. Because the Euclidean norm to the uth power is convex for $u \geq 1$, we have

$$\|\mathbb{E}_P[Z|W]\|_{P,u}^u = \mathbb{E}_P\left[\|\mathbb{E}_P[Z|W]\|^u\right] \leq \mathbb{E}_P\left[\mathbb{E}_P[\|Z\|^u|W]\right] = \mathbb{E}_P[\|Z\|^u] = \|Z\|_{P,u}^u$$

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 4.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 4.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

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

by Jensen's inequality. Therefore, we have

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

□

Lemma 4.I.9. *Consider 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 4.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 (Durrett, 1996, 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

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

□

Lemma 4.I.10. *Consider 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*

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

Proof of Lemma 4.I.10. By the Cauchy–Schwarz inequality, we have

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

Because the conditional expectation minimizes the mean squared error (Durrett, 1996, Theorem 5.1.8), we have

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

Consequently,

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

holds. □

Lemma 4.I.11. *Let $a, b \in \mathbb{R}$ be two numbers. We have*

$$(a+b)^2 \leq 2a^2 + 2b^2. \tag{4.24}$$

Proof of Lemma 4.I.11. The true statement $0 \leq (a-b)^2$ is equivalent to (4.24). □

The following lemma proved in Chernozhukov et al. (2018) states that conditional convergence in probability implies unconditional convergence in probability.

Lemma 4.I.12. *(Based on Chernozhukov et al. (2018, Lemma 6.1).) Let $\{X_t\}_{t\geq 1}$ and $\{Y_t\}_{t\geq 1}$ be sequences of random vectors and let $u \geq 1$. Consider a deterministic sequence $\{\varepsilon_t\}_{t\geq 1}$ with $\varepsilon_t \to 0$ as $t \to \infty$ such that we have $\mathbb{E}[\|X_t\|^u|Y_t] \leq \varepsilon_t^u$. Then we have $\|X_t\| = O_P(\varepsilon_t)$ unconditionally, meaning that that for any sequence $\{\ell_t\}_{t\geq 1}$ with $\ell_t \to \infty$ as $t \to \infty$ we have $P(\|X_t\| > \ell_t\varepsilon_t) \to 0$.*

Proof of Lemma 4.I.12. We have

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

by Markov's inequality. □

Lemma 4.I.13. *There exists a finite real constant C_3 satisfying $\|\beta_0\| \leq C_3$.*

Proof of Lemma 4.I.13. Recall the matrices J_0' and J_0'' in Definition 4.I.1. We have

$$\begin{aligned}\|\beta_0\| &\leq \left\|(J_0')^{-1}\right\|\left\|\mathbb{E}_P\left[A(R_X)^T\right]\right\|\left\|(J_0'')^{-1}\right\|\left\|\mathbb{E}_P\left[AR_Y\right]\right\| \\ &\leq \tfrac{1}{c_2^2}\|X\|_{P,2}\|Y\|_{P,2}\|A\|_{P,2}^2\end{aligned}$$

by submultiplicativity, Assumption 4.I.5.3, and Lemma 4.I.10. We hence infer

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

by Assumption 4.I.5.2. □

Lemma 4.I.14. *Let $\gamma \geq 0$. There exists a finite real constant C_4 satisfying $\|b^\gamma\| \leq C_4$.*

Proof of Lemma 4.I.14. We have

$$\begin{aligned}\|b^\gamma\| \;\leq\; &\left\|\left(\mathbb{E}_P\left[R_XR_X^T\right] + (\gamma-1)\,\mathbb{E}_P\left[R_XR_A^T\right]\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\mathbb{E}_P\left[R_AR_X^T\right]\right)^{-1}\right\| \\ &\cdot\left\|\mathbb{E}_P[R_XR_Y] + (\gamma-1)\,\mathbb{E}_P\left[R_XR_A^T\right]\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\mathbb{E}_P[R_AR_Y]\right\|\end{aligned}$$

by submultiplicativity. By Assumption 4.I.5.4, the largest singular value of the matrix

$$D_1 + (\gamma-1)D_2 = \mathbb{E}_P\left[R_XR_X^T\right] + (\gamma-1)\,\mathbb{E}_P\left[R_XR_A^T\right]\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\mathbb{E}_P\left[R_AR_X^T\right]$$

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

$$\|b^\gamma\| \leq \frac{1}{c_4}\Big(\|\mathbb{E}_P[R_XR_Y]\| + |\gamma-1|\left\|\mathbb{E}_P\left[R_XR_A^T\right]\right\|\left\|\mathbb{E}_P\left[R_AR_A^T\right]^{-1}\right\|\left\|\mathbb{E}_P\left[R_AR_Y^T\right]\right\|\Big)$$

by the triangle inequality and submultiplicativity. By Assumption 4.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 4.I.9 and Assumption 4.I.5.2, we have

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

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

Proof of Lemma 4.I.15. This proof is modified from Chernozhukov et al. (2018). First, 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

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

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

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

Observe that $\|U - m_U^0(W)\|_{P,2} \leq 2\|U\|_{P,2}$, and $\|V - m_V^0(W)\|_{P,2} \leq 2\|V\|_{P,2}$, and $\|Z - m_Z^0(W)\|_{P,2} \leq 2\|Z\|_{P,2}$ hold by Lemma 4.I.7. We have $\|\eta - \eta^0\|_{P,2} \leq \delta_N$ by Assumption 4.I.5.5. Therefore, we obtain the upper bound

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

by the triangle inequality, Lemma 4.I.13, Lemma 4.I.14, and Assumptions 4.I.5.2 and 4.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

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

Due to the Cauchy–Schwarz inequality, we infer

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

by Lemma 4.I.13, Lemma 4.I.14, and Assumption 4.I.5.5. Consequently, we obtain our claimed bound on λ_N. □

Lemma 4.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 4.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 4.I.16. This proof is modified from Chernozhukov et al.

(2018). By the triangle inequality, we have

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

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

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

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 $\mathcal{E}_N$ the that holds with P-probability $1 - \Delta_N$, we have

$$\begin{aligned} & \mathbb{E}_P\left[\|M_{l,t}\|^2 \middle| \{S_i\}_{i\in I_k^c}\right] \\ = & \ \tfrac{1}{n} \textstyle\sum_{i\in I_k} \mathbb{E}_P\left[|\varphi_{l,t}(S_i; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S_i; b^0, \eta^0)|^2 \middle| \{S_i\}_{i\in I_k^c}\right] \\ & \quad + \tfrac{1}{n} \textstyle\sum_{i,j\in I_k, i\neq j} \mathbb{E}_P\Big[\big(\varphi_{l,t}(S_i; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S_i; b^0, \eta^0)\big) \\ & \qquad\qquad \cdot \big(\varphi_{l,t}(S_j; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S_j; b^0, \eta^0)\big) \Big| \{S_i\}_{i\in I_k^c}\Big] \\ & \quad - 2 \textstyle\sum_{i\in I_k} \mathbb{E}_P\left[\varphi_{l,t}(S_i; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S_i; b^0, \eta^0) \middle| \{S_i\}_{i\in I_k^c}\right] \\ & \qquad\qquad \cdot \mathbb{E}_P\left[\varphi_{l,t}(S; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S; b^0, \eta^0) \middle| \{S_i\}_{i\in I_k^c}\right] \\ & \quad + n \Big| \mathbb{E}_P\left[\varphi_{l,t}(S; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S; b^0, \eta^0) \middle| \{S_i\}_{i\in I_k^c}\right] \Big|^2 \\ = & \ \mathbb{E}_P\left[|\varphi_{l,t}(S; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S; b^0, \eta^0)|^2 \middle| \{S_i\}_{i\in I_k^c}\right] \\ & \quad + \big(\tfrac{n(n-1)}{n} - 2n + n\big) \Big| \mathbb{E}_P\left[\varphi_{l,t}(S; b^0, \hat{\eta}^{I_k^c}) - \varphi_{l,t}(S; b^0, \eta^0) \middle| \{S_i\}_{i\in I_k^c}\right] \Big|^2 \\ \leq & \ \sup_{\eta\in\mathcal{T}} \mathbb{E}_P\left[\|\varphi(S; b^0, \eta) - \varphi(S; b^0, \eta^0)\|^2\right]. \end{aligned} \tag{4.25}$$

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

$$\begin{aligned} & \mathbb{E}_P\left[\|\varphi(S; b^0, \eta) - \varphi(S; b^0, \eta^0)\|^2\right] \\ \leq & \ \mathbb{E}_P[\|\varphi(S; b^0, \eta) - \varphi(S; b^0, \eta^0)\|] \\ & \quad + \mathbb{E}_P\left[\|\varphi(S; b^0, \eta) - \varphi(S; b^0, \eta^0)\|^2 \mathbb{1}_{\|\varphi(S;b^0,\eta) - \varphi(S;b^0,\eta^0)\| \geq 1}\right] \end{aligned} \tag{4.26}$$

and we have

$$\begin{aligned} &\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^2 \mathbb{1}_{\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\geq 1}\right] \\ \leq\ &\sqrt{\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^4\right]}\sqrt{P(\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\geq 1)} \end{aligned} \tag{4.27}$$

by Hölder's inequality. Observe that the term

$$\sqrt{\mathbb{E}_P\left[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|^4\right]} \tag{4.28}$$

is upper bounded by Assumption 4.I.5.5, Lemma 4.I.13 and Lemma 4.I.14. By Markov's inequality, we have

$$P(\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|\geq 1)\leq \mathbb{E}_P[\|\varphi(S;b^0,\eta)-\varphi(S;b^0,\eta^0)\|]\leq r_N^4. \tag{4.29}$$

Therefore, we have $\mathbb{E}_P[\mathcal{I}_1^2|\{S_i\}_{i\in I_k^c}]\lesssim r_N^2$ due to (4.25)–(4.29). The statistical rate r_N satisfies $r_N\lesssim \delta_N^{\frac{1}{4}}$ by Lemma 4.I.15. Thus, we infer $\mathcal{I}_1=O_P(r_N)$ by Lemma 4.I.12. Subsequently, we bound $\mathcal{I}_2$. For $r\in[0,1]$, we introduce the function

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

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

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

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 4.3.3 because the proof of Proposition 4.3.3 does neither depend on the underlying model of the random variables nor on the value of β. Furthermore, we have

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

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

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

We thus infer

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

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

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

□

Lemma 4.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 4.I.17. Consider the decomposition

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

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 $O_P(N^{-\frac{1}{2}} \rho_N)$ by Lemma 4.I.16. 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 $O_P(N^{-\frac{1}{2}})$ by the Lindeberg–Feller CLT and the Cramer–Wold device. Thus, we deduce the statement. □

Definition 4.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 4.3.1. This proof is based on Chernozhukov et al. (2018). We show the stronger statement

$$\sqrt{N}\sigma^{-1}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \overline{\psi}(S_i; \beta_0, \eta^0) + O_P(\rho_N) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d \times d}) \quad (N \to \infty), \tag{4.30}$$

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 4.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 (4.30) 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

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

We have

$$\begin{aligned}
\hat{\beta}^{\mathrm{DML2}} &= \Big(\tfrac{1}{K}\Sigma_{k=1}^{K}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\tfrac{1}{K}\Sigma_{k=1}^{K}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&= \Big(\tfrac{1}{K}\Sigma_{k=1}^{K}\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\tfrac{1}{K}\Sigma_{k=1}^{K}\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\qquad\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\qquad\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)
\end{aligned} \tag{4.31}$$

by (4.7). By Lemma 4.I.17, we have

$$\begin{aligned}
&\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
=\;& \mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)
\end{aligned} \tag{4.32}$$

and

$$\begin{aligned}
&\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
=\;& \mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big).
\end{aligned} \tag{4.33}$$

Recall the matrix J_0 introduced in Definition 4.I.1. By Weyl's inequality and Slutsky's theorem, combining Equations (4.31)–(4.33) gives

$$\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML2}}-\beta_0)\\
=&\Big(\Big(\mathbb{E}_{P_N}\Big[\big(X-m_X^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(A-m_A^0(W)\big)^T\Big]^{-1}\\
&\qquad\cdot\mathbb{E}_{P_N}\Big[\big(A-m_A^0(W)\big)\big(X-m_X^0(W)\big)^T\Big]\Big)^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\cdot \mathbb{E}_{P_N}\Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T\Big] \\
&\cdot \mathbb{E}_{P_N}\Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T\Big]^{-1} \\
&+ O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \\
\cdot \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n}}\Big(&\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&- \big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big) \\
=&\big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\big) \\
&\cdot \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n}}\Big(\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \\
&\cdot \Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big)\Big)
\end{aligned}
\tag{4.34}
$$

because K is a constant independent of N and because $N = nK$ holds. Recall the linear score ψ in (4.11). We have

$$
\sqrt{N}(\hat{\beta}^{\mathrm{DML2}} - \beta_0) = \Big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\hat{\eta}^{I_k^c}). \tag{4.35}
$$

Let $k \in [K]$. By Lemma 4.I.16, we have

$$
\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\hat{\eta}^{I_k^c}) = \frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;\beta_0,\eta^0) + O_{P_N}(\rho_N). \tag{4.36}
$$

We combine (4.35) and (4.36) 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 4.I.16, and that δ_N converges to 0 as $N \to \infty$ and that $\delta_N^{\frac{1}{4}} \geq N^{-\frac{1}{2}}$ holds by Assump-

tion 4.I.5. Thus, we have

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

We have $\mathbb{E}_P[\psi(S;\beta_0,\eta^0)]=\mathbf{0}$ due to the identifiability condition (4.5). 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 (4.30) 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

$$\sqrt{N}\sigma^{-1}(\hat{\beta}^{\mathrm{DML1}}-\beta_0)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\psi}(S_i;\beta_0,\eta^0)+O_{P_N}(\rho_N)\xrightarrow{d}\mathcal{N}(0,\mathbb{1}_{d\times d})\,(N\to\infty).$$

We have

$$\begin{aligned}
\hat{\beta}^{I_k}\;=\;&\Big(\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\cdot\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
=\;&\Big(\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\cdot\tfrac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\cdot\Big(\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\cdot\tfrac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)
\end{aligned}\tag{4.37}$$

by (4.19). Due to Weyl's inequality and Slutsky's theorem, (4.32), (4.33),

and (4.37), we obtain

$$
\begin{aligned}
&\sqrt{N}(\hat{\beta}^{\mathrm{DML1}} - \beta_0) \\
&= \Bigg(\Big(\mathbb{E}_{P_N}\Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T\Big] \\
&\qquad \cdot \mathbb{E}_{P_N}\Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T\Big]^{-1} \\
&\qquad \cdot \mathbb{E}_{P_N}\Big[\big(A - m_A^0(W)\big)\big(X - m_X^0(W)\big)^T\Big]\Big)^{-1} \\
&\quad \cdot \mathbb{E}_{P_N}\Big[\big(X - m_X^0(W)\big)\big(A - m_A^0(W)\big)^T\Big] \\
&\quad \cdot \mathbb{E}_{P_N}\Big[\big(A - m_A^0(W)\big)\big(A - m_A^0(W)\big)^T\Big]^{-1} \\
&\quad + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Bigg) \\
&\cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \Big(\tfrac{1}{\sqrt{n}}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big) \\
&\qquad\qquad - \tfrac{1}{\sqrt{n}}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big) \\
&= \Big(J_0 + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \\
&\quad \cdot \tfrac{1}{\sqrt{K}} \textstyle\sum_{k=1}^K \Big(\tfrac{1}{\sqrt{n}}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \\
&\qquad \cdot \Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\beta_0\Big)\Big).
\end{aligned}
\tag{4.38}
$$

Observe that the expression for $\sqrt{N}(\hat{\beta}^{\mathrm{DML1}} - \beta_0)$ given in (4.38) coincides with the expression for $\sqrt{N}(\hat{\beta}^{\mathrm{DML2}} - \beta_0)$ given in (4.34). 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 4.I.19. *Let $\gamma \geq 0$. Let $p > 4$ be the p from Assumption 4.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 4.I.19. Let $\eta = (m_U, m_V, m_Z) \in \mathcal{T}$. By Hölder's inequality

and the triangle inequality, we have

$$\begin{aligned}
& \mathbb{E}_P\left[\|\psi(S;b^0,\eta)\|^{\frac{p}{2}}\right]^{\frac{2}{p}} \\
= \;& \|(U-m_U(W))\big(Z-m_Z(W)-(V-m_V(W))^Tb^0\big)\|_{P,\frac{2}{p}} \\
\leq \;& \big(\|U-m_U^0(W)\|_{P,p}+\|m_U^0(W)-m_U(W)\|_{P,p}\big) \\
& \quad\cdot\big(\|Z-m_Z^0(W)\|_{P,p}+\|(V-m_V^0(W))^Tb^0\|_{P,p} \\
& \qquad +\|m_Z^0(W)-m_Z(W)\|_{P,p}+\|(m_V^0(W)-m_V(W))^Tb^0\|_{P,p}\big).
\end{aligned} \tag{4.39}$$

By the Cauchy–Schwarz inequality, we have

$$\left\|\big(V-m_V^0(W)\big)^Tb^0\right\|_{P,p} \leq \mathbb{E}_P\left[\|V-m_V^0(W)\|^p\|b^0\|^p\right]^{\frac{1}{p}} = \|b^0\|\|V-m_V^0(W)\|_{P,p} \tag{4.40}$$

and analogously

$$\left\|\big(m_V^0(W)-m_V(W)\big)^Tb^0\right\|_{P,p} \leq \|b^0\|\|m_V^0(W)-m_V(W)\|_{P,p}. \tag{4.41}$$

Hence, we infer

$$\mathbb{E}_P\left[\|\psi(S;b^0,\eta)\|^{\frac{p}{2}}\right]^{\frac{2}{p}} \leq (\|U\|_{P,p}+C_2)(\|Z\|_{P,p}+\|V\|_{P,p}+2C_2)\max\{1,\|b^0\|\} \tag{4.42}$$

by (4.39), (4.40), (4.41), Lemma 4.I.7, and Assumption 4.I.5.5. By Lemma 4.I.13, there exists a finite real constant C_3 that satisfies $\|\beta_0\| \leq C_3$. By Lemma 4.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 4.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 (4.42), we therefore have

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

□

Lemma 4.I.20. *Let $\gamma \geq 0$, and let p be as in Assumption 4.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 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 4.I.4.*

Proof of Lemma 4.I.20. This proof is modified from Chernozhukov et al. (2018). 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 (DasGupta, 2008, 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

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

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,

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

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

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

All 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 4.I.19. Thus, we have $\mathcal{I}_{k,B} = O_P(N^{\frac{p}{4}-1})$ by Lemma 4.I.12 because we have

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

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

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

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

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

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

By the Cauchy–Schwarz inequality, we have

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

All 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 4.I.19. Thus, we have

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

We hence infer $\mathcal{I}_{k,B} = O_P(N^{-\frac{1}{2}})$ by Lemma 4.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 $\max\{|b_1|,|b_2|\} \leq c$ and $\max\{|a_1-b_1|,|a_2-b_2|\} \leq d$, we have $|a_1a_2 - b_1b_2| \leq 2d(c+d)$. Indeed, we have

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

by the triangle inequality.

We apply this observation together with the triangle inequality and the Cauchy–Schwarz inequality to obtain

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

By the triangle inequality, we hence have

$$\mathcal{I}^2_{k,A} \leq 4R_{N,k}\left(\tfrac{1}{n}\textstyle\sum_{i\in I_k}\left(\big\|\psi^A_{\text{full}}(S_i)\big\|^2 + \big\|\psi^B_{\text{full}}(S_i)\big\|^2\right) + R_{N,k}\right) \tag{4.43}$$

by Lemma 4.I.11, where

$$R_{N,k} := \frac{1}{n}\sum_{i\in I_k}\left(\left\|\hat{\psi}^A_{\mathrm{full}}(S_i) - \psi^A_{\mathrm{full}}(S_i)\right\|^2 + \left\|\hat{\psi}^B_{\mathrm{full}}(S_i) - \psi^B_{\mathrm{full}}(S_i)\right\|^2\right).$$

Note that we have

$$\frac{1}{n}\sum_{i\in I_k}\left(\left\|\psi^A_{\mathrm{full}}(S_i)\right\|^2 + \left\|\psi^B_{\mathrm{full}}(S_i)\right\|^2\right) = O_P(1)$$

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 4.I.19. Thus, it suffices to bound the term $R_{N,k}$. To do this, we need to bound the four terms

$$\frac{1}{n}\sum_{i\in I_k}\|\psi(S_i;\hat{b},\hat{\eta}^{I_k^c}) - \psi(S_i;b^0,\eta^0)\|^2, \tag{4.44}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\widetilde{\psi}(S_i;\hat{b},\hat{\eta}^{I_k^c}) - \widetilde{\psi}(S_i;b^0,\eta^0)\|^2, \tag{4.45}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_1(S_i;\hat{\eta}^{I_k^c}) - \psi_1(S_i;\eta^0)\|^2, \tag{4.46}$$

$$\frac{1}{n}\sum_{i\in I_k}\|\psi_2(S_i;\hat{\eta}^{I_k^c}) - \psi_2(S_i;\eta^0)\|^2. \tag{4.47}$$

First, we bound the two terms (4.44) and (4.45) 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

$$\begin{aligned} & \tfrac{1}{n}\textstyle\sum_{i\in I_k}\|-\psi^a(S_i;\hat{\eta}^{I_k^c})(\hat{b}-b^0) + \psi(S_i;b^0,\hat{\eta}^{I_k^c}) - \psi(S_i;b^0,\eta^0)\|^2 \\ \leq\ & \tfrac{2}{n}\textstyle\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})(\hat{b}-b^0)\|^2 + \tfrac{2}{n}\textstyle\sum_{i\in I_k}\|\psi(S_i;b^0,\hat{\eta}^{I_k^c}) - \psi(S_i;b^0,\eta^0)\|^2 \end{aligned} \tag{4.48}$$

due to the triangle inequality and Lemma 4.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

$$\begin{aligned} & \tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})\|^2 \\ =\ & \tfrac{1}{n}\textstyle\sum_{i\in I_k}\left\|\big(U_i - \hat{m}_U^{I_k^c}(W_i)\big)\big(X_i - \hat{m}_X^{I_k^c}(W_i)\big)^T\right\|^2 \\ \leq\ & \sqrt{\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|U_i - \hat{m}_U^{I_k^c}(W_i)\|^4}\sqrt{\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|X_i - \hat{m}_X^{I_k^c}(W_i)\|^4} \end{aligned} \tag{4.49}$$

by the Cauchy–Schwarz inequality. We have

$$\left(\frac{1}{n}\sum_{i\in I_k}\|U_i - m_U^0(W_i)\|^4\right)^{\frac{1}{4}} = O_P(1) \tag{4.50}$$

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

$$\begin{aligned} & \mathbb{E}_P\left[\tfrac{1}{n}\textstyle\sum_{i\in I_k}\|\eta^0(W_i) - \hat{\eta}^{I_k^c}(W_i)\|^4 \middle| \{S_i\}_{i\in I_k^c}\right] \\ = \;& \mathbb{E}_P\left[\|\eta^0(W) - \hat{\eta}^{I_k^c}(W)\|^4 | \{S_i\}_{i\in I_k^c}\right] \\ \leq \;& C_2^4 \end{aligned} \tag{4.51}$$

by Assumption 4.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 4.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

$$\begin{aligned} & \tfrac{1}{n}\textstyle\sum_{i\in I_k}\|U_i - \hat{m}_U^{I_k^c}(W_i)\|^4 \\ = \;& \|U - \hat{m}_U^{I_k^c}(W)\|_{P_{I_k},4}^4 \\ \leq \;& \left(\|U - m_U^0(W)\|_{P_{I_k},4} + \|m_U^0(W) - \hat{m}_U^{I_k^c}(W)\|_{P_{I_k},4}\right)^4 \\ \leq \;& \left(\|U - m_U^0(W)\|_{P_{I_k},4} + \|\eta^0(W) - \hat{\eta}^{I_k^c}(W)\|_{P_{I_k},4}\right)^4 \\ = \;& O_P(1) \end{aligned} \tag{4.52}$$

by the triangle inequality, (4.50), and (4.51). Analogous arguments lead to

$$\frac{1}{n}\sum_{i\in I_k}\|X_i - \hat{m}_X^{I_k^c}(W_i)\|^4 = O_P(1). \tag{4.53}$$

We combine (4.49), (4.52), and (4.53) to obtain

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})\|^2 = O_P(1). \tag{4.54}$$

Because $\|\hat{b} - b^0\|^2 = O_P(N^{-1})$ holds by Theorem 4.3.1 and Theorem 4.4.1, we can bound the first summand in (4.48) by

$$\frac{1}{n}\sum_{i\in I_k}\|\psi^a(S_i;\hat{\eta}^{I_k^c})(\hat{b} - b^0)\|^2 = O_P(1)O_P(N^{-1}) = O_P(N^{-1}) \tag{4.55}$$

due to the Cauchy–Schwarz inequality and (4.54). 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 (4.48) is equal to

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

due to arguments that are analogous to (4.25)–(4.29) presented in the proof of Lemma 4.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 4.I.12. Next, we bound the two terms given in (4.46) and (4.47). We first consider the term given in (4.46). On the event $\mathcal{E}_N$, we have

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

due to arguments that are analogous to (4.25)–(4.29) presented in the proof of Lemma 4.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 4.I.12. On the event $\mathcal{E}_N$, the conditional expectation given $\{S_i\}_{i\in I_k^c}$ of the term (4.47) is given by

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

due to arguments that are analogous to (4.25)–(4.29) presented in the proof of Lemma 4.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 4.I.12. Therefore, we have $\mathcal{I}_{k,A} = O_P(N^{-\frac{1}{2}} + r_N)$ by (4.43). In total, we thus have

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

□

Theorem 4.I.21. *Suppose Assumption 4.I.5 holds. Introduce the matrix*

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

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 4.3.1, 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 4.I.4.

Proof of Theorem 4.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 4.3.1. 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

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

But this statement holds by Lemma 4.I.20 because the dimensions of A and X are fixed. □

4.J | Proofs of Section 4.4

Definition 4.J.1. *Let $\gamma \geq 0$ and recall the scalar $\rho_N = r_N + N^{\frac{1}{2}}\lambda_N$ in Definition 4.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 approximate 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 4.4.1. This proof is based on Chernozhukov et al. (2018). The matrices $D_1 + (\gamma-1)D_2$ and D_4 are invertible by Assumption 4.I.5.4. Hence, $\sigma^2(\gamma)$ is invertible.

Subsequently, we show the stronger statement

$$\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^\gamma - b^\gamma) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \overline{\psi}(S_i; b^\gamma, \eta^0) + O_P(\rho_N) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d\times d}) \,(N \to \infty), \tag{4.56}$$

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 4.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, as $N \to \infty$,

$$\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^{\gamma,\mathrm{DML2}} - b^\gamma) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \overline{\psi}(S_i; b^\gamma, \eta^0) + O_{P_N}(\rho_N) \xrightarrow{d} \mathcal{N}(0, \mathbb{1}_{d\times d}).$$

We have

$$\begin{aligned}&\hat{b}^{\gamma,\mathrm{DML2}} \\ =&\left(\frac{1}{K}\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T\big(\mathbb{1} + (\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\right)^{-1}\frac{1}{K}\sum_{k=1}^K \big(\widehat{\boldsymbol{R}}_{\boldsymbol{X}}^{I_k}\big)^T\big(\mathbb{1} + (\gamma-1)\Pi_{\widehat{\boldsymbol{R}}_{\boldsymbol{A}}^{I_k}}\big)\widehat{\boldsymbol{R}}_{\boldsymbol{Y}}^{I_k} \\ =&\left(\frac{1}{K}\sum_{k=1}^K \left(\frac{1}{n}\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\right.\right.\end{aligned}$$

$$
\begin{aligned}
&+ (\gamma-1)\cdot\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\quad\cdot\Big(\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)\Big)^{-1}\\
\cdot\frac{1}{K}\sum_{k=1}^{K}&\Big(\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&+ (\gamma-1)\cdot\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\quad\cdot\Big(\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)
\end{aligned}
\tag{4.57}
$$

by (4.14). By Lemma 4.I.17, we have

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

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

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

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

$$
\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma,\mathrm{DML2}}-b^\gamma)\\
=&\Big(\big(D_1+(\gamma-1)D_2\big)^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\cdot\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{n}}\Big(\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\\
&\qquad\cdot\Big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})-\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)b^\gamma\Big)\\
&\quad+(\gamma-1)\cdot\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)
\end{aligned}
\tag{4.58}
$$

$$
\begin{aligned}
&\cdot \Big(\frac{1}{n}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\big(\boldsymbol{A}^{I_k} - \hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T \\
&\cdot \Big(\boldsymbol{Y}^{I_k} - \hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k}) - \big(\boldsymbol{X}^{I_k} - \hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)b^\gamma\Big)\Big) \\
= &\Big(\big(D_1 + (\gamma-1)D_2\big)^{-1} + O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big) \\
&\cdot \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Big(\frac{1}{\sqrt{n}}\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \hat{\eta}^{I_k^c}) \\
&\quad + (\gamma-1)\cdot\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\cdot\big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\big)^{-1}\cdot\frac{1}{n}\sum_{i\in I_k}\psi(S_i; b^\gamma, \hat{\eta}^{I_k^c})\Big)
\end{aligned}
\tag{4.59}
$$

due to (4.57) 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 (4.58). By Lemma 4.I.16, we have

$$
\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i; b^\gamma, \hat{\eta}^{I_k^c}) \\
= \;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i; b^\gamma, \eta^0) + \Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i; b^\gamma, \hat{\eta}^{I_k^c}) - \tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i; b^\gamma, \eta^0)\Big) \\
= \;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\psi(S_i; b^\gamma, \eta^0) + O_{P_N}(\rho_N),
\end{aligned}
\tag{4.60}
$$

and

$$
\begin{aligned}
&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \hat{\eta}^{I_k^c}) \\
= \;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \eta^0) + \Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \hat{\eta}^{I_k^c}) - \tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \eta^0)\Big) \\
= \;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in I_k}\widetilde{\psi}(S_i; b^\gamma, \eta^0) + O_{P_N}(\rho_N),
\end{aligned}
\tag{4.61}
$$

and

$$
\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c}) \\
= \;&\tfrac{1}{n}\textstyle\sum_{i\in I_k}(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \psi_1(S_i;\eta^0)) + \tfrac{1}{n}\textstyle\sum_{i\in I_k}(\psi_1(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]) \\
&\quad + \mathbb{E}_{P_N}[\psi_1(S;\eta^0)] \\
= \;&O_{P_N}(N^{-\frac{1}{2}}\rho_N) + \tfrac{1}{n}\textstyle\sum_{i\in I_k}(\psi_1(S_i;\eta^0) - \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]) + \mathbb{E}_{P_N}[\psi_1(S;\eta^0)].
\end{aligned}
\tag{4.62}
$$

We apply a series expansion to obtain

$$
\begin{aligned}
&\Big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1} \\
=&\Big(\mathbb{E}_{P_N}[\psi_2(S;\eta^0)] + \frac{1}{n}\sum_{i\in I_k}\big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \psi_2(S_i;\eta^0)\big)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \eta^0) - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)] \big) \Big)^{-1} \\
&= \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \\
&\quad - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \hat{\eta}^{I_k^c}) - \psi_2(S_i; \eta^0) \big) \, \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \\
&\quad - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \eta^0) - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)] \big) \, \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \\
&\quad + O_{P_N}\Big(\Big\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \hat{\eta}^{I_k^c}) - \psi_2(S_i; \eta^0) \big) \Big\|^2 \\
&\qquad\qquad + \Big\| \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \eta^0) - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)] \big) \Big\|^2 \Big) \\
&= \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} + O_{P_N}\big(N^{-\frac{1}{2}} \rho_N \big) + O_{P_N}\Big(O_{P_N}\big(N^{-1} \rho_N^2 \big) + O_{P_N}(N^{-1}) \Big) \\
&\quad - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \eta^0) - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)] \big) \, \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \\
&= \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} + O_{P_N}\big(N^{-\frac{1}{2}} \rho_N \big) \\
&\quad - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1} \frac{1}{n} \sum_{i \in I_k} \big(\psi_2(S_i; \eta^0) - \mathbb{E}_{P_N}[\psi_2(S; \eta^0)] \big) \, \mathbb{E}_{P_N}[\psi_2(S; \eta^0)]^{-1}
\end{aligned}
\tag{4.63}
$$

due to Lemma 4.I.16, the Lindeberg–Feller CLT, the Cramer–Wold device, because $\rho_N \lesssim \delta_N^{\frac{1}{4}}$ holds by Lemma 4.I.16, and because $\delta_N^{\frac{1}{4}} \geq N^{-\frac{1}{2}}$ holds by Assumption 4.I.5. Thus, the last summand in (4.58) can be expressed as

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

$$
\begin{aligned}
&+ \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\,\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(S_i;b^\gamma,\eta^0)\\
&- \mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\,\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)\\
&\quad\cdot \mathbb{E}_{P_N}[\psi_2(S;\eta^0)]^{-1}\,\mathbb{E}_{P_N}[\psi(S;b^\gamma,\eta^0)] + O_{P_N}(\rho_N) \qquad (4.64)
\end{aligned}
$$

due to (4.60)–(4.63), the Lindeberg–Feller CLT and the Cramer–Wold device.

We combine (4.58) and (4.64) and obtain

$$
\begin{aligned}
&\sqrt{N}(\hat{b}^{\gamma,\mathrm{DML2}}-b^\gamma)\\
=\;&\Big(\big(D_1+(\gamma-1)D_2\big)^{-1}+O_{P_N}\big(N^{-\frac{1}{2}}(1+\rho_N)\big)\Big)\\
&\cdot\tfrac{1}{\sqrt{K}}\textstyle\sum_{k=1}^K\tfrac{1}{\sqrt{n}}\sum_{i\in I_k}\Big(\widetilde{\psi}(S_i;b^\gamma,\eta^0)+(\gamma-1)D_3\psi(S_i;b^\gamma,\eta^0)\\
&\quad+(\gamma-1)\big(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)D_5\\
&\quad-(\gamma-1)D_3\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)D_5\Big)+O_{P_N}(\rho_N) \qquad (4.65)\\
=\;&\Big(\big(D_1+(\gamma-1)D_2\big)^{-1}\Big)\\
&\cdot\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N\Big(\widetilde{\psi}(S_i;b^\gamma,\eta^0)+(\gamma-1)D_3\psi(S_i;b^\gamma,\eta^0)\\
&\quad+(\gamma-1)\big(\psi_1(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_1(S;\eta^0)]\big)D_5\\
&\quad-(\gamma-1)D_3\big(\psi_2(S_i;\eta^0)-\mathbb{E}_{P_N}[\psi_2(S;\eta^0)]\big)D_5\Big)+O_{P_N}(\rho_N)
\end{aligned}
$$

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 (4.56) 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, as $N\to\infty$,

$$
\sqrt{N}\sigma^{-1}(\gamma)(\hat{b}^{\gamma,\mathrm{DML1}}-b^\gamma)=\frac{1}{\sqrt{N}}\sum_{i=1}^N\overline{\psi}(S_i;b^\gamma,\eta^0)+O_{P_N}(\rho_N)\xrightarrow{d}\mathcal{N}(0,\mathbb{1}_{d\times d}).
$$

We have

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

$$
\begin{aligned}
&+ (\gamma-1)\cdot\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\quad\cdot\Big(\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\cdot\Big(\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&+ (\gamma-1)\cdot\frac{1}{n}\big(\boldsymbol{X}^{I_k}-\hat{m}_X^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\\
&\quad\cdot\Big(\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big)^{-1}\\
&\quad\cdot\frac{1}{n}\big(\boldsymbol{A}^{I_k}-\hat{m}_A^{I_k^c}(\boldsymbol{W}^{I_k})\big)^T\big(\boldsymbol{Y}^{I_k}-\hat{m}_Y^{I_k^c}(\boldsymbol{W}^{I_k})\big)\Big) \qquad (4.66)
\end{aligned}
$$

by (4.20). By Slutsky's theorem and Equation (4.66), we have

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

The last expression above coincides with 4.58. Consequently, the same asymptotic analysis conducted for $\hat{b}^{\gamma,\mathrm{DML2}}$ can also be employed in this case. □

Lemma 4.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

$$\begin{aligned} & \tfrac{1}{n}\Sigma_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)] \\ = \; & \tfrac{1}{n}\Sigma_{i\in I_k}\big(\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\big) + \tfrac{1}{n}\Sigma_{i\in I_k}\big(\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c}) - \psi(S_i;b^\gamma,\eta^0)\big) \\ & + \tfrac{1}{n}\Sigma_{i\in I_k}\big(\psi(S_i;b^\gamma,\eta^0) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\big). \end{aligned} \tag{4.67}$$

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

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

by Lemma 4.I.17. Because $\|\hat{b}^\gamma - b^\gamma\| = O_P(N^{-\frac{1}{2}}\rho_N)$ holds by Theorem 4.4.1, we infer

$$\left\|\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \frac{1}{n}\sum_{i\in I_k}\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c})\right\| = O_P\big(N^{-\frac{1}{2}}\rho_N\big). \tag{4.68}$$

Due to (4.60) that was established in the proof of Theorem 4.4.1, we have

$$\frac{1}{n}\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\hat{\eta}^{I_k^c}) - \psi(S_i;b^\gamma,\eta^0)\big) = O_P\big(N^{-\frac{1}{2}}\rho_N\big). \tag{4.69}$$

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

$$\frac{1}{n}\sum_{i\in I_k}\big(\psi(S_i;b^\gamma,\eta^0) - \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)]\big) = O_P(N^{-\frac{1}{2}}). \tag{4.70}$$

We combine (4.67) and (4.68)–(4.70) to infer the claim for $\varphi = \psi$. The case $\varphi = \widetilde{\psi}$ can be analyzed analogously. □

Theorem 4.J.3. *Suppose Assumption 4.I.5 holds. Recall the score functions introduced in Definition 4.I.1, and let* $\hat{b}^\gamma \in \{\hat{b}^{\gamma,DML1}, \hat{b}^{\gamma,DML2}\}$. *Introduce the matrices*

$$\begin{aligned} \hat{D}_1^k &:= \tfrac{1}{n}\Sigma_{i\in I_k}\psi_3(S_i;\hat{\eta}^{I_k^c}), \\ \hat{D}_2^k &:= \tfrac{1}{n}\Sigma_{i\in I_k}\psi_1(S;\hat{\eta}^{I_k^c})\Big(\tfrac{1}{n}\Sigma_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1}\tfrac{1}{n}\Sigma_{i\in I_k}\psi_1^T(S_i;\hat{\eta}^{I_k^c}), \\ \hat{D}_3^k &:= \tfrac{1}{n}\Sigma_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\Big(\tfrac{1}{n}\Sigma_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1}, \\ \hat{D}_5^k &:= \Big(\tfrac{1}{n}\Sigma_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1}\tfrac{1}{n}\Sigma_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}). \end{aligned}$$

Let furthermore

$$\begin{aligned}\widehat{\overline{\psi}}'(\cdot;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \;:=\; & \widetilde{\psi}(\cdot;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) + (\gamma-1)\hat{D}_3^k\psi(\cdot;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\ & +(\gamma-1)\big(\psi_1(\cdot;\hat{\eta}^{I_k^c}) - \tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_1(S_i;\hat{\eta}^{I_k^c})\big)\hat{D}_5^k \\ & -(\gamma-1)\hat{D}_3^k\big(\psi_2(\cdot;\hat{\eta}^{I_k^c}) - \tfrac{1}{n}\textstyle\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\big)\hat{D}_5^k\end{aligned}$$

and

$$\hat{D}_4^k := \frac{1}{n}\sum_{i\in I_k}\widehat{\overline{\psi}}'(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big(\widehat{\overline{\psi}}'(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\big)^T.$$

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

Proof of Theorem 4.J.3. This proof is based on Chernozhukov et al. (2018). We already verified

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

in the proof of Theorem 4.4.1 because K is a fixed number independent of N. Thus, we have

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

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 4.I.17.

Subsequently, we argue that $\hat{D}_5^k = D_5 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ holds. By Lemma 4.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

$$\Big(\frac{1}{n}\sum_{i\in I_k}\psi_2(S_i;\hat{\eta}^{I_k^c})\Big)^{-1} = \mathbb{E}_P[\psi_2(S;\eta^0)]^{-1} + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big). \tag{4.71}$$

Due to (4.71), it suffices to show

$$\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) = \mathbb{E}_P[\psi(S;b^\gamma,\eta^0)] + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big) \tag{4.72}$$

to infer $\hat{D}_5^k = D_5 + O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$. But (4.72) holds due to Lemma 4.J.2. To conclude the theorem, it remains verify $\hat{D}_4^k = D_4 + O_P(\tilde{\rho}_N)$. We have

$$\begin{aligned}
&\|\hat{D}_4^k - D_4\| \\
\le&\Big\|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\right]\Big\| \\
&+(\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T \\
&\quad - \mathbb{E}_P\left[\widetilde{\psi}(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]D_3^T\Big\| \\
&+(\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\quad - D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\widetilde{\psi}^T(S;b^\gamma,\eta^0)\right]\Big\| \\
&+(\gamma-1)^2\Big\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})\psi^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_3^T \\
&\quad - D_3\,\mathbb{E}_P\left[\psi(S;b^\gamma,\eta^0)\psi^T(S;b^\gamma,\eta^0)\right]D_3^T\Big\| \\
&+(\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}\widetilde{\psi}(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T \\
&\quad - \mathbb{E}_P\Big[\widetilde{\psi}(S;b^\gamma,\eta^0)D_5^T\big(\psi_1(S;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\Big]\Big\| \\
&+(\gamma-1)\Big\|\frac{1}{n}\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5\widetilde{\psi}^T(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c}) \\
&\quad - \mathbb{E}_P\Big[\big(\psi_1(S;\eta^0) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5\widetilde{\psi}^T(S;b^\gamma,\eta^0)\Big]\Big\| \\
&+(\gamma-1)^2\Big\|\frac{1}{n}\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5
\end{aligned}$$

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

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

by the triangle inequality and the results derived so far. Subsequently, we bound the terms $\mathcal{I}_1,\ldots,\mathcal{I}_{16}$ 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 4.I.20, we have

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

which implies $\mathcal{I}_1 = O_P(\tilde\rho_N)$. By Lemma 4.I.20, we have

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

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

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

and

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

By Lemma 4.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 4.I.20, we have

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

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

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

where the last summand is $O_P\big(N^{-\frac{1}{2}}(1+\rho_N)\big)$ by Lemma 4.J.2, and we have

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

The term $\mathcal{I}_6$ can be bounded analogously to $\mathcal{I}_5$. By Lemma 4.I.20, we have

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

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

$$\begin{aligned}&\Big\|\frac{1}{n}\sum_{i\in I_k}\big(\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5D_5^T\big(\psi_1(S_i;\hat{\eta}^{I_k^c})-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\\&\qquad-\mathbb{E}_P\left[\big(\psi_1(S;\eta^0)-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)D_5D_5^T\big(\psi_1(S;\eta^0)-\mathbb{E}_P[\psi_1(S;\eta^0)]\big)^T\right]\Big\|\end{aligned}$$

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

by Lemma 4.I.17, and we have

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

Next, we bound $\mathcal{I}_8$. By Lemma 4.I.20, we have

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

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

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

by Lemma 4.J.2, and we have

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

The term $\mathcal{I}_9$ can be bounded analogously to $\mathcal{I}_8$. Next, we bound $\mathcal{I}_{10}$. By Lemma 4.I.20, we have

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

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

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

by Lemma 4.J.2, and we have

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

The term $\mathcal{I}_{11}$ can be bounded analogously to $\mathcal{I}_{10}$. Next, we bound $\mathcal{I}_{12}$. By Lemma 4.I.20, we have

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

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

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

by Lemma 4.I.17, and we have

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

Next, we bound $\mathcal{I}_{13}$. By Lemma 4.I.20, we have

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

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

$$\begin{aligned}
&\Big\|\frac{1}{n}\sum_{i\in I_k}D_3\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\big(\psi_2(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)^TD_3^T \\
&\quad - D_3\,\mathbb{E}_P\Big[\psi(S;b^\gamma,\eta^0)D_5^T\big(\psi_2(S;\eta^0) - \mathbb{E}_P[\psi_2(S;\eta^0)]\big)^T\Big]D_3^T\Big\| \\
\le&\|D_3\|^2\Big\|\frac{1}{n}\sum_{i\in I_k}\psi(S_i;\hat{b}^\gamma,\hat{\eta}^{I_k^c})D_5^T\psi_2^T(S_i;\hat{\eta}^{I_k^c}) - \mathbb{E}_P\big[\psi(S;b^\gamma,\eta^0)D_5^T\psi_2^T(S;\eta^0)\big]\Big\|
\end{aligned}$$

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

by Lemma 4.J.2, and we have

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

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 4.I.20, we have

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

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

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

by Lemma 4.I.17, and we have

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

□

Proof of Proposition 4.4.2. The statement of Proposition 4.4.2 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. For instance, the Bachmann–Landau notation is presented in Lattimore and Szepesvári (2020).

Introduce the matrices

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

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 (1981, 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

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

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

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

which concludes the proof. □

Proof of Theorem 4.4.3. 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 4.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 4.4.2, 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 4.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 4.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 4.4.1. 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

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

by (4.14). Lemma 4.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 4.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

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

by (4.65). Consider the random variables

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

for $i \in [N]$, and $S_n := \sum_{i \in I_k} \widetilde{X}_i$, 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

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

holds as $n \to \infty$. Therefore, the Lindeberg–Feller condition is satisfied that 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. $\square$

Lemma 4.J.5. *Let $\gamma_N = o(\sqrt{N})$. We then have $\hat{\sigma}^2(\gamma_N) = O_P(1)$.*

Proof of Lemma 4.J.5. We have

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

As verified in the proof of Theorem 4.4.1, 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 4.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

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

in the proof of Lemma 4.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 prof of Theorem 4.J.3 because there exists some finite real constant C such that we have $|b^{\gamma_N}| \leq C$ for N large enough. Therefore,

$$\hat{\sigma}^2(\gamma_N) \;=\; \Big(\tfrac{1}{\gamma_N-1}D_1 + D_2 + o_P(1)\Big)^{-1}\tfrac{1}{(\gamma_N-1)^2}\hat{D}_4\Big(\tfrac{1}{\gamma_N-1}D_1^T + D_2^T + o_P(1)\Big)^{-1}$$

is bounded in P-probability. $\square$

Lemma 4.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 4.J.6. By Theorem 4.3.1, the term $\sqrt{N}(\beta_0 - \hat{\beta})$ asymptotically follows a Gaussian distribution and is hence bounded in P-probability. By Theorem 4.I.21, the term $\hat{\sigma}^2$ converges in P-probability. Thus, $2\hat{\sigma}$ is bounded in P-probability as well. By Lemma 4.J.4, we have $\sqrt{N}(\hat{b}^{\gamma_N} - b^{\gamma_N}) = O_P(1)$. By Lemma 4.J.5, we have $\hat{\sigma}^2(\gamma_N) = O_P(1)$. $\square$

Proof of Theorem 4.4.4. The fact the the statement holds uniformly for $P \in \mathcal{P}_N$ can be derived using analogous arguments as used to prove Theorem 4.3.1 and 4.4.1. Theorem 4.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 4.3.1. Let $\hat{\mu} := \hat{\gamma}' - 1$. We have

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

Due to Theorem 4.4.3, we have $\frac{1}{\hat{\mu}} = \frac{1}{\sqrt{N}} o_P(1)$. Due to Proposition 4.4.2, 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

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

due to Slutsky's theorem and similar arguments as presented in the proofs of Theorem 4.3.1 and 4.4.1. □

4.K | Proof of Section 4.5.1

We argue that A_1 and A_2 are independent of H conditional on W_1 and W_2 in the SEM in Figure 4.5.1. 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.

5 | Confidence and Uncertainty Assessment for Distributional Random Forests

Joint work with

Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen

This chapter is based on the manuscript

J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Inference for the distributional random forest, 2023. Preprint on arXiv:2302.05761

Abstract

The Distributional Random Forest (DRF) is a recently introduced Random Forest algorithm to estimate multivariate conditional distributions. Due to its general estimation procedure, it can be employed to estimate a wide range of targets such as conditional average treatment effects, conditional quantiles, and conditional correlations. However, only results about the consistency and convergence rate of the DRF prediction are available so far. We characterize the asymptotic distribution of DRF and develop a bootstrap approximation of it. This allows us to derive inferential tools for quantifying standard errors and the construction of confidence regions that have asymptotic coverage guarantees. In simulation studies, we empirically validate the developed theory for inference of low-dimensional targets and for testing distributional differences between two populations.

5.1 | Introduction

Building on Random Forests (Breiman, 2001), Distributional Random Forests (DRF) (Ćevid et al., 2022) provide nonparametric estimates of the distribution of a multivariate response, conditional on potentially high-dimensional covariates. DRF estimates a locally adaptive Hilbert space embedding $\hat{\mu}_n(\mathbf{x})$ of a multivariate conditional distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ of a variable of interest $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)^T \in \mathbb{R}^d$ given covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$. More precisely, in a reproducing kernel Hilbert space (RKHS) with reproducing kernel k and associated Hilbert space $\mathcal{H}$, DRF computes the estimator

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} \hat{w}_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot) \tag{5.1}$$

of the conditional mean embedding (CME) $\mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$ of $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$. The weights $\hat{w}_i(\mathbf{x})$ quantify the relevance of each training data point $\mathbf{x}_i$ to predict $\mu(\mathbf{x})$, which makes DRF locally adaptive. Ćevid et al. (2022) established consistency of $\hat{\mu}_n(\mathbf{x})$ at a fixed test point $\mathbf{x}$. A natural, but more challenging, question is whether an asymptotic normality result can be formulated for $\hat{\mu}_n(\mathbf{x})$. Providing such a result is the aim of the present paper.

We present two main results. First, we show that the appropriately centered and scaled embedding $\hat{\mu}_n(\mathbf{x})$, for a fixed test point $\mathbf{x}$, weakly converges to a limiting Gaussian process. Second, we present a resampling-based approach to infer properties of the distribution of $\hat{\mu}_n(\mathbf{x})$. In practice, this resampling-based approach allows us to simultaneously and computationally efficiently compute the DRF prediction and a bootstrap approximation of its distribution.

In addition to our theoretical developments, we present two lines of applications. First, we use the estimated Hilbert space embedding to formally test if two conditional distributions coincide or not, and we provide confidence bands for the so-called (conditional) witness function that can be used to assess where the two distributions differ. Second, we make inference for targets $\theta(\mathbf{x}) = G(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}})$ that can be represented by some smooth function G of the underlying distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ by replacing $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ by its DRF estimate. A wide range of conditional (multivariate) estimators like the conditional average treatment effect (CATE), conditional quantiles, or conditional correlations can be obtained in this way. These estimators are mutually consistent, meaning that estimated conditional covariance matrices are guaranteed to be positive semi-definite for $d < n$. In general, this might not be guaranteed if we estimated the conditional variances and covariances individually.

5.1.1 | Contributions

We develop asymptotic results for uncertainty quantification for the DRF and apply them in two use cases: testing two distributions for equality and making inference for target parameters like conditional expectations, the CATE, conditional quantiles, or conditional correlations.

We present a rigorous analysis of the DRF in an RKHS that does not depend on a specific target parameter. Consequently, the same DRF can be used to estimate different targets. Furthermore, the targets may be $\mathbb{R}^q$-valued for $q \geq 2$, and confidence ellipsoids in $\mathbb{R}^q$ can be constructed. Generalizing the arguments in Wager and Athey (2017, 2018) to RKHS's allows us to develop a U-statistics approximation of the DRF prediction $\hat{\mu}_n(\mathbf{x})$ in the RKHS. Particularly, we show that $\hat{\mu}_n(\mathbf{x})$ for a fixed test point $\mathbf{x}$ is asymptotically equivalent to a sum of independent, but not necessarily identically distributed, random elements in the Hilbert space $\mathcal{H}$. In combination with a new result on the asymptotic behavior of the variance of the DRF prediction, this result allows us to establish that

$\hat{\mu}_n(\mathbf{x})$, appropriately scaled, converges weakly to a limiting Gaussian process in the RKHS. This result holds under rather natural assumptions and does not depend on the estimation target we have in mind.

To cope with the theoretical complexity of our Hilbert space-valued Random Forest, we use and extend techniques to analyze Generalized Random Forests (GRF) (Wager and Athey, 2018; Athey et al., 2019), theory for random elements in Hilbert spaces (Hsing and Eubank, 2015; Chen and White, 1998), and bootstrap arguments (Praestgaard and Wellner, 1993; Kosorok, 2003; González-Rodríguez and Colubi, 2017). Our RKHS-valued bootstrap result builds on arguments from the bootstrap and empirical process literature and those of Athey et al. (2019). We show that an adaptation of half-sampling can be used to obtain a random element $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$ in $\mathcal{H}$, by sampling from the data, that converges to the same limiting distribution as the original estimate $\hat{\mu}_n(\mathbf{x})$, conditional on the data. Consequently, a resampling-based approach can be used to infer properties of the distribution of the random element $\hat{\mu}_n(\mathbf{x})$ of $\mathcal{H}$. In practice, we propose to adapt the DRF algorithm of Ćevid et al. (2022) to be fitted in "little bags" as motivated in Athey et al. (2019). This allows us to simultaneously and computationally efficiently compute the DRF prediction and a bootstrap approximation of its distribution in the form of $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$.

Finally, we use our bootstrap results for the DRF to formally test for distributional differences between two groups. Park et al. (2021) introduced the idea to test equality of the distributions of the control and treatment groups of an experiment, given some covariates. In contrast to estimating the CATE, which compares the two groups based on their mean, comparing whole distributions allows us to identify differences that may not be captured by the mean alone. Our developments allow us to formally test for conditional distributional differences between the control and the treatment group at a test point $\mathbf{x}$. Although it may be possible to derive an asymptotic normality result for the usual kernel-based CME estimator as used in Park et al. (2021), we are not aware of a formal test for fixed $\mathbf{x}$. Finally, our confidence bands for the conditional witness function can be interpreted as the Hilbert space-valued generalization of the work in Wager and Athey (2018); Athey et al. (2019), which derived confidence intervals for the CATE at a fixed $\mathbf{x}$.

5.1.2 | Previous Work

There is a growing literature on nonparametric estimation of multivariate conditional distributions. These include Conditional Generative Adversarial Neural Networks (Aggarwal et al., 2019), Conditional Variational Auto-Encoders (Sohn et al., 2015), Masked Autoregressive Flows (Papamakarios et al., 2017), and Conditional Mean Embeddings (Song et al., 2009; Muandet et al., 2017; Park and Muandet, 2020). To the best of our knowledge, none of these methods pro-

vide mathematical guarantees of uncertainty. Our methodology might be most closely related to the GRF, which builds on the theory of Causal Forests (Wager and Athey, 2017). GRF is a locally adaptive method to estimate univariate real-valued targets defined by local moment conditions using forest-based weights. It uses a splitting criterion for growing trees that depends on the specific estimation target, and the resulting estimator is proven to be consistent and asymptotically normal at a test point $\mathbf{x}$. In contrast to DRF, a new splitting criterion needs to be constructed for each new target and the theory presented in Athey et al. (2019) only provides results for univariate targets. However, from a theoretical perspective, GRF has exact asymptotic normality guarantees for more univariate functionals than what the current paper is able to derive with DRF because some functionals mapping $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ to the desired targets might not be sufficiently smooth. This is discussed in more detail in Remark 5.3.9. Künzel et al. (2019) introduce the X-learner to estimate the CATE, which is a meta algorithm that initially estimates the unobserved potential outcomes, and confidence intervals are obtained via the Bootstrap.

Outline: In the subsequent Section 5.2, we recall relevant definitions and results concerning RKHS's, the Landau notation, and we introduce basic concepts and summarize core ideas of the DRF. Afterwards, Section 5.3 presents our formal assumptions and main results. Section 5.4 and 5.5 discuss our two applications: inference for the conditional distributional treatment effect and general multivariate real-valued parameters. Finally, Section 5.6 demonstrates empirical validation of our theoretical developments, and Section 5.7 concludes with a brief discussion of our results.

5.2 | Background

In this section, we introduce notation and present key results from Ćevid et al. (2022) that serve as a basis for our subsequent developments. Throughout, we assume an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and denote by $\mathcal{M}_b(\mathbb{R}^d)$ the space of all bounded signed measures on $\mathbb{R}^d$.

5.2.1 | Reproducing Kernel Hilbert Spaces and Landau Notation

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the reproducing kernel Hilbert space induced by the positive definite, bounded, and continuous kernel $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$; see for instance Hsing and Eubank (2015, Chapter 2.7) for an exposition of the topic. Crucially, continuity of k ensures that $\mathcal{H}$ is separable (Hsing and Eubank, 2015, Theorem 2.7.5). For a random element ξ taking values in the (separable) Hilbert space $\mathcal{H}$ with $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$, we define its expected value in $\mathcal{H}$ by

$$\mathbb{E}[\xi] = \int_{\Omega} \xi d\mathbb{P} \in \mathcal{H},$$

where the integral is to be understood in a Bochner sense (Hsing and Eubank, 2015, Chapter 3). Because $\mathcal{H}$ is separable, this integral is well defined and there are no measurability issues. If $\mathbb{E}[\|\xi\|_{\mathcal{H}}^2] < \infty$, we define the variance of $\xi \in \mathcal{H}$ by

$$\mathrm{Var}(\xi) = \mathbb{E}[\|\xi\|_{\mathcal{H}}^2] - \|\mathbb{E}[\xi]\|_{\mathcal{H}}^2.$$

For a sequence of random elements ξ_n in $\mathcal{H}$, we denote by $\xi_n \xrightarrow{D} \xi$ convergence in distribution. That is, for all bounded and continuous functions $F\colon \mathcal{H} \to \mathbb{R}$, we have $\mathbb{E}[F(\xi_n)] \to \mathbb{E}[F(\xi)]$ as $n \to \infty$. By separability, every random element ξ with values in $\mathcal{H}$ is tight (Dudley, 2002, Chapter 7.1). That is, for all $\varepsilon > 0$, there is a compact $K_\varepsilon \subset \mathcal{H}$ such that $\mathbb{P}(\xi \in K_\varepsilon) \geq 1 - \varepsilon$. More generally, uniform tightness of a sequence ξ_n, $n \in \mathbb{N}$ means that for all $\varepsilon > 0$, there is a compact $K_\varepsilon \subset \mathcal{H}$ such that

$$\inf_n \mathbb{P}(\xi \in K_\varepsilon) \geq 1 - \varepsilon.$$

If for all $f \in \mathcal{H}$ the distribution of $\langle \xi, f \rangle$ on $\mathbb{R}$ is $N(0, \sigma_f^2)$ for some $\sigma_f > 0$, we write $\xi \sim N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ a self-adjoint Hilbert-Schmidt (HS) operator satisfying $\langle \mathbf{\Sigma} f, f \rangle = \sigma_f^2$. In this case, we also write $\xi_n \xrightarrow{D} N(0, \mathbf{\Sigma})$, if $\xi_n \xrightarrow{D} \xi$.

The kernel embedding function $\Phi\colon \mathcal{M}_b(\mathbb{R}^d) \to \mathcal{H}$ maps any bounded signed Borel measure Q on $\mathbb{R}^d$ to an element $\Phi(Q) \in \mathcal{H}$ defined by

$$\Phi(Q) = \int_{\mathbb{R}^d} k(\mathbf{y}, \cdot) dQ(\mathbf{y}) = \int_{\mathbb{R}^d} k(\mathbf{y}, \cdot) dQ^+(\mathbf{y}) - \int_{\mathbb{R}^d} k(\mathbf{y}, \cdot) dQ^-(\mathbf{y}),$$

where the integrals are Bochner integrals. Boundedness of k ensures that Φ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$. If k is the Gaussian kernel, $\|\Phi(Q_1) - \Phi(Q_2)\|_{\mathcal{H}} = 0$ implies $Q_1 = Q_2$ for all $Q_1, Q_2 \in \mathcal{M}_b(\mathbb{R}^d)$; see for example Simon-Gabriel et al. (2020, p.2) and Sriperumbudur (2016, Example 3.2). Thus, Φ is injective, and the inverse $\Phi^{-1}\colon \Phi(\mathcal{M}_b(\mathbb{R}^d)) \to \mathcal{M}_b(\mathbb{R}^d)$ is well defined. In particular, for $Q = \delta_{\mathbf{Y}}$, it holds that $\Phi(\delta_{\mathbf{Y}}) = k(\mathbf{Y}_i, \cdot)$, and thus

$$\Phi(\hat{\mathbb{P}}_{\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x}}) = \sum_{i=1}^n w_i(\mathbf{x}) \Phi(\delta_{\mathbf{Y}_i}) = \sum_{i=1}^n w_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot) = \hat{\mu}_n(\mathbf{x})$$

because Φ is linear.

For two functions f and g from the real numbers into the real numbers with $\liminf_{s \to \infty} g(s) > 0$, we write $f(s) = \mathcal{O}(g(s))$ if

$$\limsup_{s \to \infty} \frac{|f(s)|}{g(s)} \leq C$$

holds for some $0 < C < \infty$. If $C = 1$, we write $f(s) \precsim g(s)$. For a sequence of random variables $X_n\colon \Omega \to \mathbb{R}$ and $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write $X_n = \mathcal{O}_p(a_n)$

if

$$\lim_{M\to\infty} \sup_n \mathbb{P}(a_n^{-1}|X_n| > M) = 0.$$

We write $X_n = o_p(a_n)$ if $a_n^{-1}X_n$ converges in probability to zero. Similarly, for (S, d) a separable metric space, $\mathbf{X}_n \colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$, and $\mathbf{X} \colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ measurable, we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if $d(\mathbf{X}_n, \mathbf{X}) = o_p(1)$.

5.2.2 | Distributional Random Forests

Given an i.i.d. data sample of size n, DRF can be used to estimate a representation $\hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ of the conditional distribution $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)^T \in \mathbb{R}^d$ given a realization $\mathbf{x}$ of covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$ by the weighted sum

$$\hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{n} \hat{w}_i(\mathbf{x})\delta_{\mathbf{Y}_i} \tag{5.2}$$

of Dirac measures $\delta_{\mathbf{Y}_i}$. The weights $\hat{w}_i(\mathbf{x})$ quantify the relevance of a data point $\mathbf{x}_i$ in predicting the target distribution $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$.

To compute the weights $\hat{w}_i(\mathbf{x})$, DRF applies a Random Forest algorithm in the RKHS $(\mathcal{H}, k)$. That is, N trees are built, and each tree splits the data repeatedly with respect to the covariates. Each split is is chosen as to maximize the Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007) across the child nodes such that the induced distributions in the child nodes are as different as possible. For example, to split the root node of a tree, two sets of indices $\mathcal{I}_L$ and $\mathcal{I}_R$ are searched for which

$$\begin{aligned} &\left\| \Phi\left(\frac{1}{|\mathcal{I}_L|}\sum_{i\in\mathcal{I}_L}\delta_{\mathbf{Y}_i}\right) - \Phi\left(\frac{1}{|\mathcal{I}_R|}\sum_{i\in\mathcal{I}_R}\delta_{\mathbf{Y}_i}\right)\right\|_{\mathcal{H}}^2 \\ =&\left\| \frac{1}{|\mathcal{I}_L|}\sum_{i\in\mathcal{I}_L} k(\mathbf{Y}_i, \cdot) - \frac{1}{|\mathcal{I}_R|}\sum_{i\in\mathcal{I}_R} k(\mathbf{Y}_i, \cdot)\right\|_{\mathcal{H}}^2 \end{aligned} \tag{5.3}$$

is maximal. This is essentially the traditional CART splitting criterion (Breiman, 2001), but now in the RKHS. Indeed, for $d = 1$ and the kernel $k(x, y) = xy$, the MMD statistic (5.3) simplifies to the CART criterion (Ćevid et al., 2022, Section 2.3.1). Thus, the trees are built such that the distribution of the response variable in the child nodes are as different as possible in the MMD metric. Intuitively, this should lead to leaves that are as homogeneous as possible such that the leaf containing $\mathbf{x}$ of the kth tree, denoted by $\mathcal{L}_k(\mathbf{x})$, approximately contains a sample from the distribution $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$. For k being the Gaussian kernel, the embedding Φ is injective, which allows the MMD statistic to detect any distributional differences for large enough sample sizes. Crucially, this splitting criterion does not depend on the estimation target like for instance

the CATE. Ćevid et al. (2022) employed efficient computation methods of this MMD statistic to obtain a forest construction with comparable computational complexity as the original Random Forest algorithm. This is achieved by using a well-known approximation of the MMD statistic with a specified number of random features (Ćevid et al., 2022, Section 2.3).

Once the trees are grown and leaf nodes determined, the weights $w_i(\mathbf{x})$ can be computed. For each tree $k = 1, \ldots, N$, the leaf node $\mathcal{L}_k(\mathbf{x})$ of the kth tree is the leaf in which $\mathbf{x}$ falls. Then, the prediction of $\mu(\mathbf{x})$ from each tree is given by averaging the elements $k(\mathbf{Y}_j, \cdot)$ that belong to $\mathcal{L}_k(\mathbf{x})$, namely $1/|\mathcal{L}_k(\mathbf{x})| \sum_{j \in \mathcal{L}_k(\mathbf{x})} k(\mathbf{Y}_j, \cdot)$; that is, the $k(\mathbf{Y}_j, \cdot)$'s belonging to the leaf $\mathcal{L}_k(\mathbf{x})$ of $\mathbf{x}$ each get assigned the weight $1/|\mathcal{L}_k(\mathbf{x})|$. These per-tree predictions are subsequently averaged to form the forest predictor

$$\hat{\mu}_n(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{|\mathcal{L}_k(\mathbf{x})|} \sum_{j \in \mathcal{L}_k(\mathbf{x})} k(\mathbf{Y}_j, \cdot) \right).$$

Rearranging this double sum such that each Hilbert element is present only once yields

$$\hat{\mu}_n(\mathbf{x}) = \sum_{i=1}^{n} \hat{w}_i(\mathbf{x}) k(\mathbf{Y}_i, \cdot)$$

for suitable weights $\hat{w}_i(\mathbf{x})$. From this last expression, we can read off our weights $\hat{w}_i(\mathbf{x})$ that quantify the importance of the ith data point in predicting $\mu(\mathbf{x}) = \Phi(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}})$. Consequently, this approach allows us to characterize data-adaptive neighborhoods of data points $\mathbf{x}$ whose corresponding conditional distribution is similar to $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$.

5.3 | Theoretical Development

DRF estimates the embedding $\mu(\mathbf{x}) = \Phi(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]$ of the conditional distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ in an RKHS with reproducing kernel k. In this section, we first state the assumptions on the forest construction and the data generating process and recall that it consistently estimates $\mu(\mathbf{x})$ at a certain rate (Ćevid et al., 2022). Subsequently, we establish convergence in distribution of the standardized estimator to a limiting Gaussian process. Lastly, we develop a consistent variance estimation procedure that enables efficient empirical computation.

5.3.1 | Forest Construction and Consistency in the RKHS

We require our forest construction to satisfy the following properties that are similar to Wager and Athey (2018). First, we require that the data used to build a tree is independent from the data used to populate its leaves for prediction. To ensure this, we split the subsample used to build a particular tree into two

halves. The first half is used to construct the tree. Then, the data from the second half gets assigned to the leaves of the tree according to the covariate splits that were fitted on the first half. Subsequently, the responses from the second half of the data, which are now distributed across the leaves, are used to form the DRF predictions. Second, when a parent node is split into two child nodes, every feature may be chosen with at least a certain non-zero probability. Third, the prediction of a tree is not allowed to depend on the order of the training samples. Fourth, when a parent node of a tree is split into two child nodes, this split may not be arbitrarily imbalanced. Each child node needs to contain a certain fraction α of its parent's data points. Finally, to grow a tree, the traditional Random Forest algorithm samples training data points with replacement from the n training points; that is, a bootstrap approach is pursued. In contrast, we sample a subset without replacement as done by Wager and Athey (2018); Athey et al. (2019). These assumptions on the forest construction are summarized as follows:

(**F1**) (*Honesty*) The data used for constructing each tree is split into two halves; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response.

(**F2**) (*Random-split*) At every split point and for all feature dimensions $j = 1, \ldots, p$, the probability that the split occurs along the feature X_j is bounded from below by π/p for some $\pi > 0$.

(**F3**) (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

(**F4**) (α-*regularity*) After splitting a parent node, each child node contains at least a fraction $\alpha \leq 0.2$ of the parent's training samples. Moreover, the trees are grown until every leaf contains between κ and $2\kappa - 1$ many observations for some fixed tuning parameter $\kappa \in \mathbb{N}$.

(**F5**) (*Data sampling*) To grow a tree, a subsample of size s_n out of the n training data points is sampled. We consider $s_n = n^\beta$ with

$$1 > \beta > \left(1 + \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p}\right)^{-1},$$

where α is chosen in (**F4**).

The validity of the above properties are ensured by the forest construction. As outlined above, the prediction of DRF for a given test point $\mathbf{x}$ is an element of $\mathcal{H}$. If we denote the ith training observation by $\mathbf{Z}_i = (\mathbf{X}_i, k(\mathbf{Y}_i, \cdot)) \in \mathbb{R}^p \times \mathcal{H}$,

then DRF estimates the embedding of the conditional distribution $\Phi(\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}})$ by averaging the corresponding estimates across the N trees, namely

$$\Phi(\hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}})) = \frac{1}{N}\sum_{k=1}^{N} T(\mathbf{x};\varepsilon_k,\mathcal{Z}_k),$$

where $\mathcal{Z}_k = \{\mathbf{Z}_{k_1},\ldots,\mathbf{Z}_{k_{s_n}}\}$ is a random subset of $\{\mathbf{Z}_i\}_{i=1}^n$ of size s_n (see (**F5**)) chosen for constructing the kth tree, and ε_k is a random variable capturing the randomness in growing the kth tree such as the choice of the splitting candidates, and $T(\mathbf{x};\varepsilon_k,\mathcal{Z}_k)$ denotes the output of a single tree. The output of a single tree is given by the average of the terms $k(\mathbf{Y}_i,\cdot)$ over all data points $\mathbf{X}_i$ contained in the leaf $\mathcal{L}_k(\mathbf{x})$ of the tree constructed from ε_k and $\mathcal{Z}_k$:

$$T(\mathbf{x};\varepsilon_k,\{\mathbf{Z}_{k_1},\ldots,\mathbf{Z}_{k_{s_n}}\}) = \sum_{j=1}^{s_n}\frac{\mathbb{1}(\mathbf{X}_{k_j}\in\mathcal{L}_k(\mathbf{x}))}{|\mathcal{L}_k(\mathbf{x})|}k(\mathbf{Y}_{k_j},\cdot). \tag{5.4}$$

To develop our theory, we do not consider forests that consist of a user-specified number N of trees. Instead, we consider $N\to\infty$, such that the forest estimator $\hat{\mu}_n(\mathbf{x})$ is obtained by averaging all possible $\binom{n}{s_n}$ many trees, which equals the number of possible subsets of $\{\mathbf{Z}_i\}_{i=1}^n$ of size s_n. This idealized version of our DRF predictor, which we will denote by $\hat{\mu}_n(\mathbf{x})$ from now onwards, is given by

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1}\sum_{i_1<i_2<\cdots<i_{s_n}}\mathbb{E}_\varepsilon\left[T(\mathbf{x};\varepsilon,\{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_{s_n}}\})\right]. \tag{5.5}$$

This is a standard simplification also employed by Wager and Athey (2017, 2018); Athey et al. (2019). Ćevid et al. (2022) established that $\hat{\mu}_n(\mathbf{x})$ in (5.5) consistently estimates $\mu(\mathbf{x})$ with respect to the RKHS norm at a certain rate.

Theorem 5.3.1 (Theorem 1 in Ćevid et al. (2022))**.** *Assume that the forest construction satisfies the properties (**F1**)–(**F5**). Additionally, assume that k is a bounded and continuous kernel (this corresponds to Assumption (**K1**) and (**K2**) below) and that we have a random design with $\mathbf{X}_1,\ldots,\mathbf{X}_n$ independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity (this corresponds to (**D1**) below). If the subsample size s_n is of order n^β for some $0<\beta<1$, the mapping $\mathbf{x}\mapsto\mu(\mathbf{x})\in\mathcal{H}$ is Lipschitz (this corresponds to (**D2**) below) and*

$$\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2\mid\mathbf{X}=\mathbf{x}]<\infty$$

*(this is a consequence of assumption (**D3**) below). Then, we have consis-*

tency of $\hat{\mu}_n(\mathbf{x})$ *in* (5.5) *with respect to the RKHS norm, namely*

$$\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right) \tag{5.6}$$

for any $\gamma \leq \frac{1}{2}\min\left(1-\beta, \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}\cdot\beta\right)$.

Although this result shows consistency of $\hat{\mu}_n(\mathbf{x})$ at a certain rate, it does not establish distributional convergence of the scaled difference. Subsequently, we establish this result.

5.3.2 | Asymptotic Normality in the RKHS

To establish an asymptotic Gaussian process behavior of $\hat{\mu}_n(\mathbf{x})$ in the Hilbert space, we first show asymptotic linearity in Theorem 5.3.2. More precisely, we show that

$$\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) = \frac{s_n}{n}\sum_{i=1}^{n} T_n(\mathbf{Z}_i) + o_p(\sigma_n)$$

holds, where $\mathbf{Z}_i = (\mathbf{X}_i, k(\mathbf{Y}_i, \cdot)) \in \mathbb{R}^d \times \mathcal{H}$ concatenates the ith covariates and the embedding of the ith response in the Hilbert space, T_n is some function depending on n, and σ_n is some standard deviation converging to zero. Denote by

$$\xi_n = \frac{1}{\sigma_n}\left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right) \tag{5.7}$$

the shifted and scaled DRF estimator whose asymptotic distribution we subsequently investigate. To establish that ξ_n asymptotically converges to a Gaussian process, two ingredients are required (Hsing and Eubank, 2015, Chapter 7). First, we require weak convergence to a limiting Gaussian distribution in $\mathbb{R}$ of the univariate marginals $\langle \frac{s_n}{n\sigma_n}\sum_{i=1}^{n} T_n(\mathbf{Z}_i), f\rangle$ for all $f \in \mathcal{H}$. Second, we require uniform tightness of the sequence $\frac{s_n}{n\sigma_n}\sum_{i=1}^{n} T_n(\mathbf{Z}_i)$ for $n \geq 1$.

We make the following assumptions on the data generating process. Throughout, we assume all involved expectations exist and are finite.

(**D1**) The covariates $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed on $[0,1]^p$ with a density bounded away from 0 and infinity.

(**D2**) The mapping $\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}$ is Lipschitz.

(**D3**) The mapping $\mathbf{x} \mapsto \mathbb{E}[\|k(\mathbf{Y}, \cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]$ is Lipschitz.

(**D4**) $\mathrm{Var}(k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[\|k(\mathbf{Y}, \cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] - \|\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 > 0$.

(**D5**) $\mathbb{E}[\|k(\mathbf{Y},\cdot)-\mu(\mathbf{x})\|_{\mathcal{H}}^{2+\delta} \mid \mathbf{X}=\mathbf{x}] \leq M$, for some constants δ, M uniformly over $\mathbf{x} \in [0,1]^d$.

(**D6**) For all $f \in \mathcal{H}$, $\mathrm{Var}(\langle k(\mathbf{Y},\cdot), f\rangle_{\mathcal{H}} \mid \mathbf{X}=\mathbf{x}) = \mathrm{Var}(f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}) > 0$.

(**D7**) For all $f \in \mathcal{H}$, $\mathbf{x} \mapsto \mathbb{E}[|f(\mathbf{Y})|^2 \mid \mathbf{X}=\mathbf{x}]$ is Lipschitz.

Assumption (**D1**) is a standard assumption when analyzing Random Forests (Meinshausen, 2006; Wager and Athey, 2017, 2018), and (**D2**)–(**D5**) correspond to natural generalizations of the assumptions in Wager and Athey (2018) to the RKHS setting. Particularly, Assumption (**D2**) implies that we have

$$\|\mu(\mathbf{x}_1)-\mu(\mathbf{x}_2)\|_{\mathcal{H}} \leq L\|\mathbf{x}_1-\mathbf{x}_2\|$$

for some $L>0$. Because $\|\cdot\|_{\mathcal{H}}$ metrizes weak convergence for the Gaussian kernel, the distributions $\mathbb{P}_{\mathbf{Y}\mid\mathbf{X}=\mathbf{x}_1}$ and $\mathbb{P}_{\mathbf{Y}\mid\mathbf{X}=\mathbf{x}_2}$ are consequently close in the weak topology if $\mathbf{x}_1$ and $\mathbf{x}_2$ are close enough in $\mathbb{R}^p$. Moreover, (**D2**) implies that for all $f \in \mathcal{H}$ and all $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$, we have

$$\begin{aligned}|\mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}_1] - \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}_2]| &= |\langle f, \mu(\mathbf{x}_1)-\mu(\mathbf{x}_2)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}}\|\mu(\mathbf{x}_1)-\mu(\mathbf{x}_2)\|_{\mathcal{H}} \\ &\leq \|f\|_{\mathcal{H}} L\|\mathbf{x}_1-\mathbf{x}_2\|. \end{aligned} \tag{5.8}$$

Consequently, (**D2**) implies Lipschitz continuity of $\mathbf{x} \mapsto \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]$ for all $f \in \mathcal{H}$. Similarly, (**D5**) implies that for all $f \in \mathcal{H}$,

$$\begin{aligned}&\mathbb{E}[|f(\mathbf{Y}) - \mathbb{E}[f(\mathbf{Y}) \mid \mathbf{X}=\mathbf{x}]|^{2+\delta} \mid \mathbf{X}=\mathbf{x}] \\ &= \mathbb{E}[|\langle f, k(\mathbf{Y},\cdot)-\mu(\mathbf{x})\rangle|^{2+\delta} \mid \mathbf{X}=\mathbf{x}] \\ &\leq \mathbb{E}[|\|f\|_{\mathcal{H}} \cdot \|k(\mathbf{Y},\cdot)-\mu(\mathbf{x})\|_{\mathcal{H}}|^{2+\delta} \mid \mathbf{X}=\mathbf{x}] \\ &\leq \|f\|_{\mathcal{H}}^{2+\delta} M \end{aligned} \tag{5.9}$$

holds uniformly over $\mathbf{x} \in [0,1]^d$. These two conclusions together with **(D6)** and **(D7)** will allow us to apply results of Wager and Athey (2018) for the univariate marginal

$$\left\langle \frac{s_n}{n\sigma_n} \sum_{i=1}^{n} T_n(\mathbf{Z}_i), f \right\rangle = \frac{s_n}{n\sigma_n} \sum_{i=1}^{n} \langle T_n(\mathbf{Z}_i), f\rangle$$

to establish the asymptotic normality of these marginals. Finally, **(D1)** and **(D3)** imply that for any $\mathbf{x} \in [0,1]^d$ and some $\mathbf{x}_0 \in [0,1]^d$,

$\mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}]$

$$
\begin{aligned}
&= \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] - \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}_0] + \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}_0] \\
&\leq L\|\mathbf{x}-\mathbf{x}_0\| + \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}_0] \\
&\leq 2L\sqrt{p} + \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}_0],
\end{aligned}
$$

so that

$$
\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}}^2 \mid \mathbf{X}=\mathbf{x}] < \infty, \tag{5.10}
$$

as required in Theorem 5.3.1 above.

We also make the following assumptions on the kernel k:

(**K1**) k is bounded.

(**K2**) $(\mathbf{x},\mathbf{y}) \mapsto k(\mathbf{x},\mathbf{y})$ is (jointly) continuous.

(**K3**) k is integrally strictly positive definite (denoted by $\int$spd), that is

$$
\|Q_1 - Q_2\|_{\mathcal{H}} = 0 \implies Q_1 = Q_2, \text{ for all } Q_1, Q_2 \in \mathcal{M}_b(\mathbb{R}^d);
$$

see for instance Sriperumbudur (2016); Simon-Gabriel et al. (2020).

The Gaussian kernel satisfies the conditions in (**K1**)–(**K3**), for instance.

As outlined above, our first main result shows that ξ_n in (5.7) is asymptotically linear, that is, indistinguishable from a sum of independent elements in $\mathcal{H}$ as $n \to \infty$.

Theorem 5.3.2. *Assume conditions (**F1**)–(**F5**), (**D1**)–(**D7**), (**K1**), and (**K2**) hold. Denote by $\mathbf{Z}_i = (\mathbf{X}_i, k(\mathbf{Y}_i,\cdot))$, $i = 1,\ldots,n$. Then, there exists a map $T_n\colon [0,1]^p \times \mathcal{H} \to \mathcal{H}$ such that, with*

$$
\sigma_n^2 = \frac{s_n^2}{n}\mathrm{Var}(T_n(\mathbf{Z}_1)), \tag{5.11}
$$

we have $\sigma_n \to 0$, $\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\| = \mathcal{O}_p(\sigma_n)$, and

$$
\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) = \frac{s_n}{n}\sum_{i=1}^{n} T_n(\mathbf{Z}_i) + o_p(\sigma_n). \tag{5.12}
$$

Moreover, T_n is given by

$$
T_n(\mathbf{Z}_i) = \mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)]. \tag{5.13}
$$

Remark 5.3.3. *To decrease the bias of the individual trees, the subsample size s_n must not be of too small order compared to n. However, this causes*

the variance σ_n^2 *to go to* 0 *at a slower rate than* $\sqrt{n}$*, and the precise rate is given by*

$$C_1 \frac{\sqrt{s_n}}{\log(s_n)^{d/2}\sqrt{n}} \precsim \sigma_n \precsim C_2 \frac{\sqrt{s_n}}{\sqrt{n}}$$

similarly to Wager and Athey (2018). If $s_n = n^\beta$ *with* β *as in* ***(F5)****, this translates to*

$$C_1 \frac{1}{\beta^{d/2}\log(n)^{d/2} n^{(1-\beta)/2}} \precsim \sigma_n \precsim C_2 \frac{1}{n^{(1-\beta)/2}}.$$

Due to Theorem 5.3.2, it is enough to show that

$$\frac{s_n}{n\sigma_n} \sum_{i=1}^{n} T_n(\mathbf{Z}_i) \xrightarrow{D} N(0, \mathbf{\Sigma_x})$$

to establish asymptotic normality of ξ_n. To achieve this, we need to establish univariate convergence and asymptotic tightness.

For $f \in \mathcal{H}$, consider the univariate marginal $\frac{s_n}{n}\sum_{i=1}^{n}\langle T_n(\mathbf{Z}_i), f\rangle$. Due to Assumption (**F1**)–(**F5**) and Lipschitz continuity of $\mathbf{x} \mapsto \langle \mu(\mathbf{x}), f\rangle$ implied by (**D1**) and (5.9), Assumption (**D1**)–(**D7**) verify all assumptions of Theorem 3.1 of Wager and Athey (2018). Consequently, there exists a $\sigma_n(f) > 0$ converging to zero with n such that

$$\left\langle \frac{1}{\sigma_n(f)} \left(\frac{s_n}{n} \sum_{i=1}^{n} T_n(\mathbf{Z}_i) \right), f \right\rangle_{\mathcal{H}} \xrightarrow{D} N(0,1). \tag{5.14}$$

Unfortunately, the scaling factor $\sigma_n(f)$ obtained from Wager and Athey (2018) depends on f. The challenge is to show that the convergence in (5.14) holds for any $f \in \mathcal{H}$ if $\sigma_n(f)$ is replaced by σ_n given in (5.11). To establish this, we need to refine the characterization of the asymptotic behavior of the variance of T_n. The following result achieves this.

Theorem 5.3.4. *Assume conditions* ***(F1)****–****(F5)****,* ***(D1)****–****(D7)****,* ***(K1)****, and* ***(K2)*** *hold. Then, for all* $f \in \mathcal{H} \setminus \{0\}$*, we have*

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(T_n(\mathbf{Z}_1))} = \frac{\mathrm{Var}(\langle k(\mathbf{Y}, \cdot), f\rangle | \mathbf{X} = \mathbf{x})}{\mathrm{Var}(k(\mathbf{Y}, \cdot)|\mathbf{X} = \mathbf{x})} = \sigma^2(f) > 0. \tag{5.15}$$

Thus, the variance of the first order approximation of the univariate forest prediction is of the same order as that of the forest prediction in the Hilbert space. That the resulting ratio $\sigma^2(f)$ is strictly larger than zero is a consequence of assumption (**D6**).

The convergence in (5.14) together with Theorem 5.3.4 establishes

$$\left\langle \frac{s_n}{n\sigma_n} \sum_{i=1}^{n} T_n(\mathbf{Z}_i), f \right\rangle_{\mathcal{H}} \xrightarrow{D} N(0, \sigma^2(f)),$$

that is, weak convergence of the univariate marginals $\langle \frac{s_n}{n\sigma_n} \sum_{i=1}^{n} T_n(\mathbf{Z}_i), f \rangle$ for all $f \in \mathcal{H}$. Establishing additionally uniform tightness (Hsing and Eubank, 2015, Chapter 7) yields our second main result, namely the asymptotic Gaussian process distribution of the DRF prediction.

Theorem 5.3.5. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then,*

$$\frac{1}{\sigma_n} \left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) \right) \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{\mathbf{x}}), \tag{5.16}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}$ is a self-adjoint HS operator satisfying

$$\langle \boldsymbol{\Sigma}_{\mathbf{x}} f, f \rangle = \frac{\mathrm{Var}(\langle k(\mathbf{Y}, \cdot), f \rangle | \mathbf{X} = \mathbf{x})}{\mathrm{Var}(k(\mathbf{Y}, \cdot) | \mathbf{X} = \mathbf{x})} > 0 \tag{5.17}$$

for all $f \in \mathcal{H}$.

The expression of $\boldsymbol{\Sigma}_{\mathbf{x}}$ is intuitive: if $\boldsymbol{\Sigma}_{\mathbf{x}}^{o}$ is the covariance operator of the random element $k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}$, then $\boldsymbol{\Sigma}_{\mathbf{x}}$ equals $\boldsymbol{\Sigma}_{\mathbf{x}}^{o}$ standardized by its trace; see for example Hsing and Eubank (2015, Chapter 7).

We now turn to the question of how to approximate the distribution of $\hat{\mu}_n(\mathbf{x})$ itself.

5.3.3 | Approximation of the Sampling Distribution

In this section, we establish an approach to approximate the sampling distribution of ξ_n based on half-sampling. This can afterwards be used to make inference for derived point estimators or functionals.

Our half-sampling scheme is motivated by Athey et al. (2019) and is as follows. For a subset $\mathcal{S} \subset \{1, \ldots, n\}$ with $s_n \leq |\mathcal{S}|$, denote by $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$ the version of $\hat{\mu}_n(\mathbf{x})$ that only uses trees built with data from $\mathcal{S}$. That is, $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$ is the counterpart of $\boldsymbol{\Phi}_{\mathcal{H}}$ in Athey et al. (2019). In Athey et al. (2019), $\mathcal{S}$ was randomly drawn without replacement such that $|\mathcal{S}| = n/2$. To simplify our theoretical developments in approximating the whole distribution of $\hat{\mu}_n(\mathbf{x})$, we draw $\mathcal{S}$ by sampling n i.i.d. random variables $W_i \sim \mathrm{Bernoulli}(1/2)$ and consider $\mathcal{S} = \{i \colon W_i = 1\}$. The cardinality $|\mathcal{S}|$ of $\mathcal{S}$ randomly fluctuates around $n/2$, with $|\mathcal{S}|/n \to 1/2$ almost surely. Because $\mathcal{S}$ is chosen at random, the element $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})$ now has two sources of randomness: one from the data and one from drawing $\mathcal{S}$. Subsequently, we establish that, if the data are kept fixed and only

the randomness of the choice of $\mathcal{S}$ is considered,

$$\xi_n^{\mathcal{S}} = \frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right) \tag{5.18}$$

converges to the same Gaussian random element as the original process ξ_n. This allows us to estimate the whole distribution and characteristic quantities such as variances from its subsample versions by randomly drawing $\mathcal{S}$.

To establish this result, we build on standard bootstrap arguments as for instance presented in Kosorok (2008, Chapter 10). Formally, we establish in Theorem 5.3.6 that

$$\xi_n^{\mathcal{S}} = \frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right) \xrightarrow[W]{D} N(0, \mathbf{\Sigma}_{\mathbf{x}}) \tag{5.19}$$

holds. The symbol $\xrightarrow[W]{D}$ denotes so-called conditional convergence in distribution and is characterized by the condition

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}[h(\xi_n^{\mathcal{S}}) \mid \mathcal{Z}_n] - \mathbb{E}[h(\xi)]\right| \xrightarrow{p} 0, \tag{5.20}$$

where $\mathrm{BL}_1(\mathcal{H})$ denotes the space of all bounded Lipschitz functions from $\mathcal{H}$ to $\mathbb{R}$ with Lipschitz constant bounded by 1. That is, $h \in \mathrm{BL}_1(\mathcal{H})$ satisfies $\sup_{f \in \mathcal{H}} |h(f)| \leq 1$ and $|h(f_1) - h(f_2)| \leq \|f_1 - f_2\|_{\mathcal{H}}$ for all $f_1, f_2 \in \mathcal{H}$. This definition is in particular reasonable if we recall that convergence in distribution alone, $\xi_n \xrightarrow{D} \xi$, is characterized by $\sup_{h \in \mathrm{BL}_1(\mathcal{H})} |\mathbb{E}[h(\xi_n)] - \mathbb{E}[h(\xi)]| \to 0$; see for example Dudley (2002, Theorem 11.3.3). Consequently, (5.20) means that, conditional on the data $\mathcal{Z}_n$, $\xi_n^{\mathcal{S}}$ converges to ξ in distribution in probability; see for example González-Rodríguez and Colubi (2017); Kosorok (2008, Chapter 10). Hence, if condition (5.20) holds, we write (5.19).

Combining arguments from Kosorok (2003); González-Rodríguez and Colubi (2017) with those from Athey et al. (2019), we show that:

Theorem 5.3.6. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then,* (5.19) *holds.*

Consequently, for "large" n, the distribution of $\xi_n^{\mathcal{S}}$, given the data, is the same as that of ξ_n. To empirically characterize this distribution, we use a similar approximation trick as in Athey et al. (2019). We grow our forest by (i) drawing B subsets $\mathcal{S}_1, \ldots, \mathcal{S}_B$ of $\{1, \ldots, n\}$ as described above, (ii) fitting a DRF with ℓ trees and calculating the prediction $\hat{\mu}_n^{\mathcal{S}_b}(\mathbf{x})$ for each $b = 1, \ldots, B$, and (iii) obtaining the overall prediction $\hat{\mu}_n(\mathbf{x})$ as the average over $(\hat{\mu}_n^{\mathcal{S}_b}(\mathbf{x}))_{b=1}^B$. This allows us to obtain both an overall DRF prediction and B i.i.d. draws from the distribution of $\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right)$. This can then be used to approximate, for

instance, the variance of $F(\hat{\mu}_n(\mathbf{x}))$ for some function F. The following result establishes consistency of this approach for linear and continuous $F\colon \mathcal{H} \to \mathbb{R}^q$.

Corollary 5.3.7. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then, for any $F\colon \mathcal{H} \to \mathbb{R}^q$ linear and continuous,*

$$\mathbb{E}\left[\frac{1}{\sigma_n^2}\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)^\top \,\middle|\, \mathcal{Z}_n\right] \xrightarrow{p} F \circ \mathbf{\Sigma}_{\mathbf{x}}. \tag{5.21}$$

This in particular implies the result for F appropriately differentiable. Crucially, it is also possible to estimate σ_n itself.

Corollary 5.3.8. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then,*

$$\frac{\mathbb{E}[\|\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n]}{\sigma_n^2} \xrightarrow{p} 1. \tag{5.22}$$

Remark 5.3.9. *The class of suitable differentiable functions $F\colon \mathcal{H} \to \mathbb{R}^q$ depends on the chosen kernel k. We will focus on the Gaussian kernel. This has several advantages: the Gaussian kernel meets all assumptions **(K1)**–**(K3)** and metrizes weak convergence. Thus, the convergence in $\mathcal{H}$ in* (5.6) *can be interpreted as convergence of $\hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ to $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ in the weak topology. Moreover, the Gaussian kernel can be computationally efficiently approximated with the techniques in Ćevid et al. (2022). However, the RKHS induced by the Gaussian kernel is a relatively small space of functions. For instance, for $d = 1$, the identity function $f(y) = y$ is not contained in $\mathcal{H}$ for the Gaussian kernel (Minh, 2010, Theorem 3). Thus, if we desire to estimate the conditional mean of Y with $\hat{\mu}_n(\mathbf{x})$, asymptotic normality is not automatically guaranteed by our result. However, because $\mathcal{H}$ is dense in the space of bounded and continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$ (Minh, 2010), it is conceivable that the asymptotic normality result extends to this case. Indeed, our simulation results in Section 5.6 indicate that approximate normality holds for a wide range of functionals, and crucially also for functions into $\mathbb{R}^q$ for $q > 1$.*

5.4 | Application: Conditional Distributional Treatment Effect

A frequent measure to assess the effectiveness of a binary treatment W given some covariates $\mathbf{X} = \mathbf{x}$ is the CATE, $\mathbb{E}[Y^{do(W=1)} - Y^{do(W=0)} \mid \mathbf{X} = \mathbf{x}]$, where we use the do-notation of Pearl (1995).

As in Park et al. (2021), we assume that strong ignorability holds, that is, (i) unconfoundedness $W \perp\!\!\!\perp (\mathbf{Y}^0, \mathbf{Y}^1) \,|\, \mathbf{X}$ and (ii) overlap $0 < \mathbb{P}(W = 1 \,|\, \mathbf{X}) < 1$. In this case, the CATE can be estimated as a difference in estimated conditional expectations at $\mathbf{X} = \mathbf{x}$. That is, the expected mean difference between the treatment and control groups among subjects with properties $\mathbf{x}$ is considered. Although the CATE allows us to take treatment effect heterogeneity into account due to conditioning on the covariates $\mathbf{X}$, it fails to capture distributional differences between the treatment and control groups beyond the mean. The conditional distributional treatment effect (CoDiTE) (Park et al., 2021) alleviates this problem. The idea of CoDiTE (with the conditional mean embedding) is to not only compare expected values of the treatment and control groups, but to extend the comparison to more general aspects of the distributions. To achieve this, a kernel estimator of the conditional mean embedding, CME, is used (Song et al., 2009, 2013; Park and Muandet, 2020). For instance, to test whether there are any distributional differences between the treatment and the control groups, CME's of both groups are computed and compared. The kernel method of Park et al. (2021) requires choosing two kernels and does not come with formal hypothesis testing. In contrast, we can estimate the CME's of the two groups by two DRF's in a locally adaptive way instead of choosing a kernel for the covariate space. Moreover, we are able to introduce tests and confidence bands at a test point $\mathbf{x}$ using the Gaussian Hilbert space element approximation we derived above.

Let us denote by $\hat{\mu}_{n_0,0}(\mathbf{x})$ the DRF estimate in the control group ($W = 0$) and by $\hat{\mu}_{n_1,1}(\mathbf{x})$ the estimate in the treatment group ($W = 1$), and let $\mathbb{P}^0_{\mathbf{Y} \,|\, \mathbf{X}=\mathbf{x}}$ and $\mathbb{P}^1_{\mathbf{Y} \,|\, \mathbf{X}=\mathbf{x}}$ be the associated conditional distributions of the control and treatment groups at the test point $\mathbf{x}$, respectively. The conditional witness function (Park et al., 2021)

$$\mathbb{R}^d \ni \mathbf{y} \mapsto \hat{\mu}_{n_1,1}(\mathbf{x})(\mathbf{y}) - \hat{\mu}_{n_0,0}(\mathbf{x})(\mathbf{y}) \in \mathcal{H} \tag{5.23}$$

is a means to capture differences between the two conditional distributions $\mathbb{P}^0_{\mathbf{Y} \,|\, \mathbf{X}=\mathbf{x}}$ and $\mathbb{P}^1_{\mathbf{Y} \,|\, \mathbf{X}=\mathbf{x}}$ as a function of the response value $\mathbf{y}$. The true conditional witness function is given by

$$\mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}) = \mathbb{E}[k(\mathbf{Y}^1, \mathbf{y}) \,|\, \mathbf{X} = \mathbf{x}] - \mathbb{E}[k(\mathbf{Y}^0, \mathbf{y}) \,|\, \mathbf{X} = \mathbf{x}].$$

Areas of $\mathbf{y}$-values where the conditional witness function is positive or negative indicate where the conditional density of one group is higher or lower than the other (Park et al., 2021). If the conditional witness function is non-zero, there are distributional differences between the treatment and the control group. Such a comparison is especially helpful if the conditional mean estimates in the

two groups are equal, resulting in a conditional treatment effect of 0 on the mean level.

Our developments in this section are as follows. First, we present a formal test for assessing whether the conditional response distributions of the treatment and control groups are equal. Particularly, we develop a test for

$$H_0 \colon \mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \quad \text{vs.} \quad H_A \colon \mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \neq \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \tag{5.24}$$

using the statistic $\|\hat{\mu}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})\|^2_{\mathcal{H}}$, which equals the norm of the conditional witness function in the Hilbert space. We establish that our test is asymptotically valid and, given a ∫spd kernel as in (**K3**), the power of our test converges to 1. Second, we provide a simultaneous asymptotic confidence band for the conditional witness function itself. These two developments involve the distribution of the squared norm of the Gaussian random element $\|\xi\|^2_{\mathcal{H}}$, which is intractable (Gretton et al., 2012). Our half-sampling approach presents a convenient way to approximate this distribution.

Before we present our results, we introduce some notation. Denote by n_0 the size of the control group and by n_1 the size of the treatment group. For simplicity, we assume that $n_0/n_1 \to 1$, but it is possible to relax this condition. Let $(\mathbf{Y}^0_i, \mathbf{X}_i)$, $i = 1, \ldots, n_0$ and $(\mathbf{Y}^1_i, \mathbf{X}_i)$, $i = 1, \ldots, n_1$ denote i.i.d. samples from the control and treatment groups, respectively, and let $\mathcal{Z}^j_{n_j} = \{(k(\mathbf{Y}^j_1, \cdot), \mathbf{X}_1), \ldots, (k(\mathbf{Y}^j_{n_j}, \cdot), \mathbf{X}_{n_j})\}$ for $j \in \{0, 1\}$ denote the respective observations with response elements of the Hilbert space $\mathcal{H}$. We denote the concatenated data from both groups by $\mathcal{Z}_{n_{01}} = (\mathcal{Z}_{n_0}, \mathcal{Z}_{n_1})$, and introduce the total number of observations $n_{01} = n_0 + n_1$. We assume that the observations from the treatment and control groups are independent and that strong ignorability holds as in Park et al. (2021). Furthermore, let $\xi_j \sim N(0, \mathbf{\Sigma}^j_{\mathbf{x}})$ for $j \in \{0, 1\}$, where for all $f \in \mathcal{H}$

$$\langle \mathbf{\Sigma}^j_{\mathbf{x}} f, f \rangle = \frac{\mathrm{Var}(\langle k(\mathbf{Y}^j, \cdot), f \rangle | \mathbf{X} = \mathbf{x})}{\mathrm{Var}(k(\mathbf{Y}^j, \cdot) | \mathbf{X} = \mathbf{x})} \tag{5.25}$$

holds as in Theorem 5.3.5 with the respective variance-covariance operators from both groups. Finally, let $\sigma_{n_j,j}$ denote the standard deviation as in (5.11) for the respective groups $j \in \{0, 1\}$.

The following result describes the asymptotic distribution of the (suitably rescaled) test statistic for the testing problem (5.24). Moreover, the result establishes that the same limiting distribution is obtained if the individual "subforests" of the DRF are used as a bootstrap sample, as described in Section 5.3.3. This will allow us to approximate the distribution of the test statistic for testing (5.24) and for formulating a simultaneous confidence band for the conditional witness function.

Corollary 5.4.1. *Assume conditions **(F1)**–**(F5)** and **(D1)**–**(D7)** for both groups, **(K1)**, and **(K2)** hold, together with strong ignorability. Also assume that $n_0, n_1 \to \infty$ with $n_0/n_1 \to 1$. Then, for $\mathcal{S}_0$, $\mathcal{S}_1$ independent,*

$$\left\| \frac{1}{\sigma_{n_1,1}} (\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}} (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow[W]{D} \|\xi_0 - \xi_1\|^2_{\mathcal{H}} \tag{5.26}$$

and

$$\left\| \frac{1}{\sigma_{n_1,1}} (\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}} (\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow{D} \|\xi_0 - \xi_1\|^2_{\mathcal{H}}. \tag{5.27}$$

Moreover, if the ratio $\sigma_{n_0,0}/\sigma_{n_1,1}$ converges to some real number $c_2(\mathbf{x})$ that is bounded away from 0 and ∞ as the sample sizes n_0, n_1 tend to infinity, we obtain

$$\frac{1}{\sigma^2_{n_1,1}} \left\| (\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow[W]{D} \|\xi_0 - c_2(\mathbf{x})\xi_1\|^2_{\mathcal{H}} \tag{5.28}$$

and

$$\frac{1}{\sigma^2_{n_1,1}} \left\| (\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - (\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow{D} \|\xi_0 - c_2(\mathbf{x})\xi_1\|^2_{\mathcal{H}}. \tag{5.29}$$

The above result assumes convergence of the ratio $\sigma_{n_0,0}/\sigma_{n_1,1}$. This condition is used to obtain a common scaling factor in (5.28) and (5.29). With the expressions derived in Theorem 5.3.4, it reduces to assuming

$$\frac{\mathrm{Var}(\mathbb{E}[\frac{1}{N^0_{\mathbf{x}}} \mathbb{1}\{\mathbf{X}_2 \in \mathcal{L}^0(\mathbf{x})\} \mid \mathbf{X}_1])}{\mathrm{Var}(\mathbb{E}[\frac{1}{N^1_{\mathbf{x}}} \mathbb{1}\{\mathbf{X}_2 \in \mathcal{L}^1(\mathbf{x})\} \mid \mathbf{X}_1])} \to c(\mathbf{x}) > 0. \tag{5.30}$$

This essentially means that the behavior of the (conditional) variance of the respective leaf node is asymptotically of the same order in both samples. Given the assumptions on the forest, together with strong ignorability, this seems to be a mild condition. The common scaling factor and limiting behavior in (5.28) and (5.29) allows us to use a bootstrap procedure on the "subforests" to approximate the distribution of the test statistic to test (5.24). The convergence in (5.26) and (5.28) should be understood conditional on the joint data $\mathcal{Z}_{n_{01}}$ from both groups.

Subsequently, we describe how Corollary 5.4.1 can be used to formally test the

hypothesis (5.24). In particular, we explain how to approximate the distribution of our test statistic $\sigma_{n_1,1}^{-2}\|\hat{\mu}_{n_0,0}(\mathbf{x})-\hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2$ under the null hypothesis. Under the null $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$, we have $\mu_0(\mathbf{x}) = \mu_1(\mathbf{x})$. Consequently, (5.29) describes the asymptotic distribution of the rescaled test statistic, namely

$$\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 \xrightarrow{D} \|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2. \tag{5.31}$$

Thus, the rescaled test statistic $\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2$ has the same limiting distribution as its resampling bootstrap version

$$\frac{1}{\sigma_{n_1,1}^2}\left\|(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x}))\right\|_{\mathcal{H}}^2 \tag{5.32}$$

given the data. Moreover, we can (approximately) obtain this distribution by sampling from $\mathcal{S}$, irrespective of whether $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ holds. Hence, the distribution of the rescaled test statistic $\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2$ under the null hypothesis can be obtained by sampling from $\mathcal{S}$. Particularly, let $c_{n_1,\alpha}$ be the smallest value obtained from B such draws with B sufficiently large such that

$$\mathbb{P}\left(\frac{1}{\sigma_{n_1,1}^2}\left\|(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x}))\right\|_{\mathcal{H}}^2 > c_{n_1,\alpha} \;\middle|\; \mathcal{Z}_{n_{01}}\right) \leq \alpha \tag{5.33}$$

holds. That is, $c_{n_1,\alpha}$ is the $1-\alpha$ quantile of the test statistic simulated under the null. Next, we establish that the same number $c_{n_1,\alpha}$ can be used to formulate a corresponding test for the test statistic computed on the full data. Define the test $\varphi(\mathcal{Z}_{n_{01}})$ for our testing problem by

$$\varphi(\mathcal{Z}_{n_{01}}) = \mathbb{1}\left\{\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n_1,\alpha}\right\}.$$

The following result establishes that φ is of level α and that its power converges to 1.

Theorem 5.4.2. *Assume conditions **(F1)**–**(F5)** and **(D1)**–**(D7)** for both groups, **(K1)**–**(K3)** hold, together with strong ignorability and* (5.30)*. Then, as $n_0, n_1 \to \infty$ such that $n_0/n_1 \to 1$,*

(i) φ has a valid type-I error. That is, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$,

$$\limsup_{n_0,n_1} \mathbb{P}\left(\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n_1,\alpha}\right) \leq \alpha.$$

(ii) φ has power going to 1. That is, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \neq \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$,

$$\lim_{n_0,n_1} \mathbb{P}\left(\frac{1}{\sigma^2_{n_1,1}} \|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|^2_{\mathcal{H}} > c_{n_1,\alpha}\right) = 1.$$

In practice, the scaling factor $1/\sigma^2_{n_1,1}$ is unknown. In principle, it can be estimated as elaborated in Corollary 5.3.8. However, we can directly consider the unscaled resampled statistics (5.32), namely $\|(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x}))\|^2_{\mathcal{H}}$, and identify its $1-\alpha$ quantile, which corresponds to $\sigma^2_{n_1,1}c_{n_1,\alpha}$.

Subsequently, we present a procedure to construct a confidence band for the conditional witness function $\mathbf{y} \mapsto \mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y})$ that is valid jointly for all $\mathbf{y}$-values. Let $c_{n_1,\alpha}$ be as in (5.33). We show in the following theorem that the interval

$$\begin{aligned}\mathcal{B}(\mathbf{y}) = [&\hat{\mu}_{n_1,1}(\mathbf{x})(\mathbf{y}) - \hat{\mu}_{n_0,0}(\mathbf{x})(\mathbf{y}) - \sqrt{c_{n_1,\alpha}C}\sigma_{n_1,1},\\ &\hat{\mu}_{n_1,1}(\mathbf{x})(\mathbf{y}) - \hat{\mu}_{n_0,0}(\mathbf{x})(\mathbf{y}) + \sqrt{c_{n_1,\alpha}C}\sigma_{n_1,1}]\end{aligned} \tag{5.34}$$

is a $1-\alpha$ confidence band for the conditional witness function, where $C = \sup_{\mathbf{y}} k(\mathbf{y},\mathbf{y})$. The constant C is finite due to assuming boundedness of the reproducing kernel in Assumption (**K2**). That is, $\mathcal{B}(\mathbf{y})$ is a confidence band for the conditional witness function that is valid jointly for all $\mathbf{y}$.

Theorem 5.4.3. *Assume conditions (**F1**)–(**F5**) and (**D1**)–(**D7**) for the control and the treatment group, and assume that (**K1**) and (**K2**) hold together with strong ignorability and* (5.30)*. Then, for $\mathcal{B}(\mathbf{y})$ as in* (5.34)*, with $n_0, n_1 \to \infty$ such that $n_0/n_1 \to 1$,*

$$\liminf_{n_0,n_1\to\infty} \mathbb{P}\left(\cap_{\mathbf{y}}\{\mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}) \in \mathcal{B}(\mathbf{y})\}\right) \geq 1-\alpha. \tag{5.35}$$

Similarly to above, when performing finite sample calculations and $\sigma_{n_1,1}$ is unknown, we can estimate $\sqrt{c_{n_1,\alpha}}\sigma_{n_1,1}$ using the same resampling procedure as above. Furthermore, we have $C = 1$ if we use the Gaussian kernel.

5.4.1 | Computation

Subsequently, we provide details on the computation of the test statistic $\|\hat{\mu}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})\|^2_{\mathcal{H}}$ for testing equality of the distributions of the control and the treatment group as well as the confidence band for the conditional witness function.

Consider the three real-valued kernel matrices $\mathbf{K}_0 = (k(\mathbf{Y}^0_i, \mathbf{Y}^0_j))_{i=1,\dots n_0, j=1,\dots n_0}$ and $\mathbf{K}_1 = (k(\mathbf{Y}^1_i, \mathbf{Y}^1_j))_{i=1,\dots n_1, j=1,\dots n_1}$ and $\mathbf{K} = (k(\mathbf{Y}^0_i, \mathbf{Y}^1_j))_{i=1,\dots n_0, j=1,\dots, n_1}$. Denote by $\hat{\mathbf{w}}_0 \in \mathbb{R}^{n_0}$ and $\hat{\mathbf{w}}_1 \in \mathbb{R}^{n_1}$ the vectors that concatenate the weights from the DRF predictors for the control and treatment groups, respectively.

Moreover, for $j \in \{0, 1\}$, consider $\mathbf{k}_j = (k(\mathbf{Y}_1^j, \cdot), \ldots, k(\mathbf{Y}_{n_j}^j, \cdot))^\top$, and denote by $\mathbf{k}_j(\mathbf{y}) = (k(\mathbf{Y}_1^j, \mathbf{y}), \ldots, k(\mathbf{Y}_{n_j}^j, \mathbf{y}))^\top$ for $\mathbf{y} \in \mathbb{R}^d$. Then, we have

$$\begin{aligned}
\hat{\mu}_{n_0,0}(\mathbf{x}) &= \sum_{i=1}^{n_0} \hat{w}_{i,0}(\mathbf{x}) k(\mathbf{Y}_i^0, \cdot) = \hat{\mathbf{w}}_0^\top \mathbf{k}_0, \\
\hat{\mu}_{n_1,1}(\mathbf{x}) &= \sum_{i=1}^{n_1} \hat{w}_{i,1}(\mathbf{x}) k(\mathbf{Y}_i^1, \cdot) = \hat{\mathbf{w}}_1^\top \mathbf{k}_1, \\
\hat{\mu}_{n_0,0}^{\mathcal{S}_0}(\mathbf{x}) &= \sum_{i=1}^{n_0} \hat{w}_{i,0}^{\mathcal{S}_0}(\mathbf{x}) k(\mathbf{Y}_i^0, \cdot) = (\hat{\mathbf{w}}_0^{\mathcal{S}_0})^\top \mathbf{k}_0, \\
\hat{\mu}_{n_1,1}^{\mathcal{S}_1}(\mathbf{x}) &= \sum_{i=1}^{n_1} \hat{w}_{i,1}^{\mathcal{S}_1}(\mathbf{x}) k(\mathbf{Y}_i^1, \cdot) = (\hat{\mathbf{w}}_1^{\mathcal{S}_1})^\top \mathbf{k}_1.
\end{aligned}$$

Subsequently, we compute $c_{n_1,\alpha}\sigma_{n_1,1}^2$ as the $1-\alpha$ quantile of the B many draws from

$$\begin{aligned}
&\|\hat{\mu}_{n_0,0}^{\mathcal{S}_0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x}) - (\hat{\mu}_{n_1,1}^{\mathcal{S}_1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x}))\|_{\mathcal{H}}^2 \\
=&(\hat{\mathbf{w}}_0^{\mathcal{S}_0} - \hat{\mathbf{w}}_0)^\top \mathbf{K}_0 (\hat{\mathbf{w}}_0^{\mathcal{S}_0} - \hat{\mathbf{w}}_0) + (\hat{\mathbf{w}}_1^{\mathcal{S}_1} - \hat{\mathbf{w}}_1)^\top \mathbf{K}_1 (\hat{\mathbf{w}}_1^{\mathcal{S}_1} - \hat{\mathbf{w}}_1) \\
&- 2(\hat{\mathbf{w}}_0^{\mathcal{S}_0} - \hat{\mathbf{w}}_0)^\top \mathbf{K} (\hat{\mathbf{w}}_1^{\mathcal{S}_1} - \hat{\mathbf{w}}_1).
\end{aligned}$$

To test the null hypothesis of having an equal distribution in the control and the treatment group according to (5.24), we first compute the test statistic $\|\hat{\mu}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})\|_{\mathcal{H}}^2$ according to

$$\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 = \hat{\mathbf{w}}_0^\top \mathbf{K}_0 \hat{\mathbf{w}}_0 + \hat{\mathbf{w}}_1^\top \mathbf{K}_1 \hat{\mathbf{w}}_1 - 2\hat{\mathbf{w}}_0^\top \mathbf{K} \hat{\mathbf{w}}_1. \tag{5.36}$$

The confidence band for the conditional witness function is then given by

$$\mathcal{B}(\mathbf{y}) = [\hat{\mathbf{w}}_1^\top \mathbf{k}_1(\mathbf{y}) - \hat{\mathbf{w}}_0^\top \mathbf{k}_0(\mathbf{y}) - \sqrt{c_{n_1,\alpha} C}, \hat{\mathbf{w}}_1^\top \mathbf{k}_1(\mathbf{y}) - \hat{\mathbf{w}}_0^\top \mathbf{k}_0(\mathbf{y}) + \sqrt{c_{n_1,\alpha} C}], \tag{5.37}$$

where we have $C = 1$ for the Gaussian kernel.

5.5 | Application: General Real-Valued Parameters

The asymptotic normality result for $\hat{\mu}_n(\mathbf{x})$ derived in Section 5.3 can also be applied to make inference for q-dimensional real-valued parameters $\theta(\mathbf{x})$ that can be expressed as a function G of the underlying conditional distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$, namely $\theta(\mathbf{x}) = G(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}})$. The DRF predictor estimates the embedding $\hat{\mu}_n(\mathbf{x})$ of $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ in the Hilbert space. This embedding can then be "pulled back" to give an estimator $\hat{\mathbb{P}}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ of $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ that can in turn be used to estimate $\theta(\mathbf{x})$. More precisely, we can represent our estimator $\hat{\theta}(\mathbf{x})$ by

$\hat{\theta}(\mathbf{x}) = F(\hat{\mu}_n(\mathbf{x}))$ for some function F that maps from the Hilbert space into $\mathbb{R}^q$. For sufficiently smooth F, the asymptotic normality of $\frac{1}{\sigma_n}(\hat{\theta}(\mathbf{x}) - \theta(\mathbf{x}))$ follows from Theorem 5.3.5.

In practice, we estimate $\theta(\mathbf{x})$ by $\hat{\theta}(\mathbf{x}) = G(\hat{\mathbb{P}}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}})$, where $\hat{\mathbb{P}}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ is the "pull-back" of the DRF predictor $\hat{\mu}_n(\mathbf{x})$ as in (5.2). To compute confidence intervals for the individual components of $\theta(\mathbf{x})$, we first compute subsample estimators $\hat{\theta}^{\mathcal{S}_b}(\mathbf{x}) = G(\hat{\mathbb{P}}^{\mathcal{S}_b}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}})$ for $b = 1, \ldots, B$, where $\hat{\mathbb{P}}^{\mathcal{S}_b}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ corresponds to the pullback of the subsample DRF predictor $\hat{\mu}^{\mathcal{S}_b}(\mathbf{x})$. Then, the empirical variance $\widehat{\mathrm{Var}}(\hat{\theta}(\mathbf{x}))$ of $\hat{\theta}(\mathbf{x})$ can be estimated by the variance of the $\hat{\theta}^{\mathcal{S}_b}(\mathbf{x})$ over $b = 1, \ldots, B$, and confidence intervals can be built using the Gaussian approximation. Alternatively, it is possible to compute confidence intervals via the approximate sampling distribution. To pursue this approach, one first computes the $1 - \alpha/2$ quantile $\hat{q}_{1-\alpha/2}$ and and the $\alpha/2$ quantile $\hat{q}_{\alpha/2}$ of $\{\hat{\theta}^{\mathcal{S}_b}(\mathbf{x}) - \hat{\theta}(\mathbf{x})\}_{b=1\ldots,B}$. Component-wise $1-\alpha$ confidence intervals for two-sided testing of $\theta(\mathbf{x}) = \mathbf{0}$ are then given by $[\hat{\theta}(\mathbf{x}) - \hat{q}_{1-\alpha/2}, \hat{\theta}(\mathbf{x}) - \hat{q}_{\alpha/2}]$.

For multi-dimensional parameters $\theta(\mathbf{x})$, which corresponds to $q > 1$, one can compute simultaneous elliptical confidence balls. If we denote the $q \times q$ covariance matrix obtained from the sample $\{\hat{\theta}^{\mathcal{S}_b}(\mathbf{x}) - \hat{\theta}(\mathbf{x})\}_{b=1\ldots,B}$ by $\widehat{\mathrm{Var}}(\hat{\theta}(\mathbf{x}))$, these consist of all parameters τ such that the resulting test statistic $||\widehat{\mathrm{Var}}(\hat{\theta}(\mathbf{x}))^{-1/2}(\hat{\theta}(\mathbf{x}) - \tau)||^2$ is smaller than the $1 - \alpha$ quantile of a $\chi^2(q)$ distribution with q degrees of freedom. Analogously to above, one may use the approximate sampling distribution of $||\widehat{\mathrm{Var}}(\hat{\theta}(\mathbf{x}))^{-1/2}(\hat{\theta}^{\mathcal{S}_b}(\mathbf{x}) - \hat{\theta}(\mathbf{x}))||^2$ instead of the $\chi^2(q)$ distribution.

5.6 | Empirical Results

In this section, we demonstrate the performance of our DRF confidence intervals for the CATE, conditional quantiles, conditional correlations, and conditional witness functions for simulated data. We consider almost exclusively data generating mechanisms that have already been considered by Ćevid et al. (2022). The only adaptation is that we consider $U(-1, 1)^p$ distributed covariates $\mathbf{X}$ instead of $U(0, 1)^p$ in Section 5.6.3. In all examples except for the conditional witness functions, we grow a forest that consists of $B = 100$ subforests with $\ell = 1000$ trees each, and we choose $\beta = 0.9$ in assumption (**F5**). To fit trees, 10 random features are used for the approximation of the MMD statistic when splitting the nodes, and the minimal node size is 5. Moreover, we consider the Gaussian kernel with the median bandwidth heuristic and compute confidence intervals using the Gaussian approximation. For the conditional witness functions, we consider forests that consist of $B = 200$ subforests with $\ell = 1000$ trees each and choose $\beta = 0.9$ because estimating whole confidence bands for the conditional witness function is a complicated task. Code of our analysis

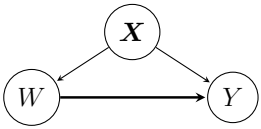


Figure 5.6.1: Causal graph illustrating the data generating processes in (5.38) and (5.39).

is available on GitHub (`https://github.com/JeffNaef/drfinference`).

We demonstrate that DRF performs well for a wide range of estimation targets $\theta(\mathbf{x})$. The effort of the user is minimal because estimating a DRF does not depend on the actual target(s).

5.6.1 | Conditional Average Treatment Effect

Subsequently, we perform inference for CATE's between a control group $W = 0$ and a treatment group $W = 1$. We thereby follow the approach used in Ćevid et al. (2022) and consider W as a part of the response, using DRF to find the conditional distribution of $(Y, W) \mid \mathbf{X} = \mathbf{x}$. This agrees with our view of seeing the (causal) parameter of interest as a function F of the CME $\hat{\mu}_n(\mathbf{x})$ and, under strong ignorability, consistency of this approach follows from the consistency of $\hat{\mu}_n(\mathbf{x})$. This approach is different from Wager and Athey (2018); Athey et al. (2019) who consider W as a part of the covariates.

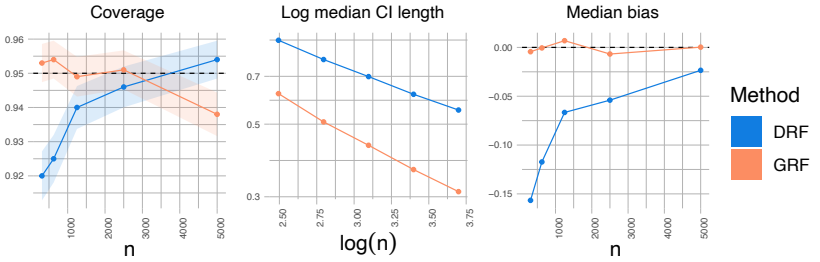
First, we consider a situation where the treatment effect is homogeneous but where Y and W are confounded by X_3. We simulate data from

$$\begin{aligned} &\mathbf{X} \sim \text{Unif}(0,1)^5, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\big(0.25(1+\beta_{2,4}(X_3))\big) \\ &Y \mid (\mathbf{X}, W) \sim 2(X_3 - 0.5) + \mathcal{N}(0,1), \end{aligned} \tag{5.38}$$

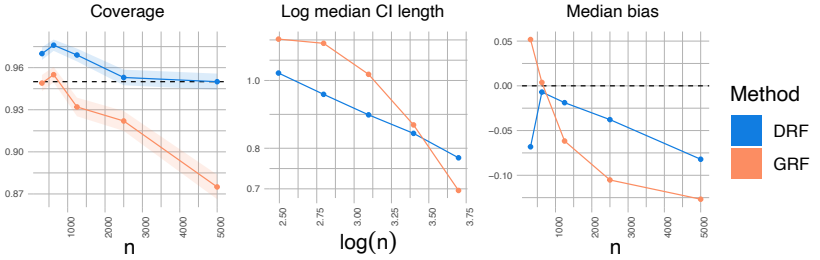
where $\beta_{a,b}$ denotes the density of a beta-distributed random variable with parameters a and b. We consider the test point $\mathbf{x} = (0.7, 0.3, 0.5, 0.68, 0.43)^T$. Our results and comparisons to GRF obtained over 1000 simulation runs are displayed in Figure 5.6.2a. The performance of DRF improves as the sample size increases and it reaches the nominal coverage level. GRF undercovers for $n = 5000$. However, GRF outperforms DRF with respect to coverage for small sample sizes due to its small bias in this example. Moreover, the confidence intervals of GRF are shorter than the ones with DRF.

Second, we consider a situation where the treatment effect is heterogeneous and where Y and W are confounded. We simulate data from

$$\begin{aligned} &\mathbf{X} \sim \text{Unif}(0,1)^5, \quad W \mid \mathbf{X} \sim \text{Bernoulli}\big(0.25(1+\beta_{2,4}(X_3))\big) \\ &Y \mid (\mathbf{X}, W) \sim 2(X_3 - 0.5) + (W - 0.2) \cdot \eta(X_1)\eta(X_2) + \mathcal{N}(0,1), \end{aligned} \tag{5.39}$$



(a) Without treatment effect



(b) With treatment effect

Figure 5.6.2: Estimating the CATE of Y given $\mathbf{X} = \mathbf{x} = (0.7, 0.3, 0.5, 0.68, 0.43)^T$ with data from (5.38) (homogeneous treatment effect and observed confounding) in Figure 5.6.2a and with data from (5.39) (heterogeneous treatment effect and observed confounding) in Figure 5.6.2b for different values of n over 1000 simulation runs. The plots display the coverage (fraction of times the true, and in general unknown, CATE was inside the confidence interval) and log median length of two-sided 95% confidence intervals for the CATE and median bias over 1000 simulation runs. The shaded regions in the coverage plots represent 95% confidence bands with respect to the 1000 simulation runs. DRF parameters: $B = 100$, $\ell = 1000$, $\beta = 0.9$, consider 10 randomly sampled features to split, minimal node size of 5. GRF parameters: 50 000 trees, other values are left at their default values.

where $\eta(x) = 1 + (1 + \exp -20(x - 1/3))^{-1}$ and $\beta_{a,b}$ denotes the density of a beta-distributed random variable with parameters a and b. That is, the treatment effect is heterogeneous because different values of $\mathbf{X}$ result in a different treatment effect, and confounding via $\mathbf{X}$ is present because W also depends on $\mathbf{X}$. We consider the test point $\mathbf{x} = (0.7, 0.3, 0.5, 0.68, 0.43)^T$. Our results and comparisons to GRF obtained over 1000 simulation runs are displayed in Figure 5.6.2b. For small sample sizes n, DRF overcovers, but it gradually reaches the nominal 95% level for larger sample sizes. In contrast, GRF fails to reach the nominal 95% level for larger sample sizes due to its bias.

When estimating the CATE with the GRF algorithm, a centering step to center Y and W with respect to $\mathbf{X}$ is performed. With DRF, we found that such an additional centering is not useful. With DRF, we used a total number of 10^5 trees whereas with GRF, we were not able to use as many due to computational reasons. Since the `drf` package (Michel and Ćevid, 2021) used is based on `grf` (Tibshirani et al., 2022), this indicates empirically that the target-tailored splitting criterion of GRF can be computationally considerably more expensive than the general splitting criterion of DRF.

5.6.2 | Conditional Quantiles

Subsequently, we consider performing inference for conditional quantiles of $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$. We consider simulated data where the response variable Y experiences a shift in its mean depending on the value of X_1, namely

$$\mathbf{X} \sim \text{Unif}(-1, 1)^5, \quad Y \sim \mathcal{N}\big(0.8 \cdot \mathbb{1}_{X_1>0}, 1\big). \tag{5.40}$$

The results for estimating three conditional quantiles (10%, 50%, and 90%), a sample size of $n = 5000$, and a range of x_1-values are displayed in Figure 5.6.3 and 5.6.4. In Figure 5.6.3, the coverage for the different quantiles is close to the nominal 95% coverage except at and around the value $x_1 = 0$ where the mean function of Y experiences a discontinuity. Figure 5.6.4 displays the joint coverage of all three conditional quantiles 10%, 50%, and 90%. The coverage is again close to the nominal and slightly higher than it for x_1-values away from 0. The disturbing effect of the discontinuity at $x_1 = 0$ is again visible.

5.6.3 | Conditional Correlation

Conditional copulas allow us to represent conditional multivariate distributions $\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y_1 \leq y_1, \ldots, Y_d \leq y_d \mid \mathbf{X} = \mathbf{x})$ in terms of the marginal distributions $\mathbb{P}(Y_i \leq y \mid \mathbf{X} = \mathbf{x}) = F_{Y_i \mid \mathbf{X}=\mathbf{x}}(y)$ for $1 \leq i \leq d$. This technique is frequently employed in fields such as risk analysis or finance (Cherubini et al., 2004). More precisely, Sklar's theorem (Sklar, 1959) asserts the existence of a so-called conditional copula $C_{\mathbf{x}}$ at the test point $\mathbf{x}$, which is a

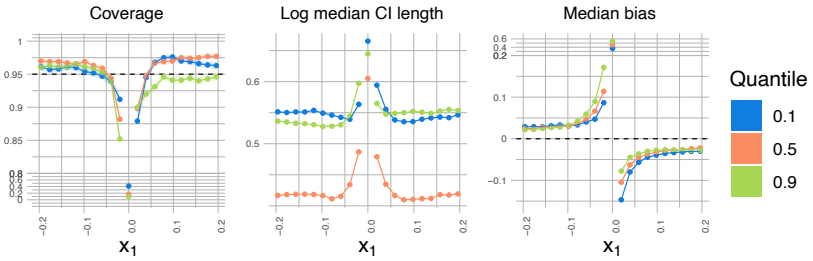


Figure 5.6.3: Estimating conditional quantiles (10%, 50%, and 90%; differentiated by color) of Y given $X_1 = x_1$ from (5.40) (mean shift in Y based on X_1) for $n = 5000$ and different values of x_1. The plot displays the coverage (fraction of times the true, and in general unknown, conditional quantile was inside the confidence interval) and log median length of two-sided 95% confidence intervals for the conditional quantile and median bias over 1000 simulation runs. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1000 simulation runs. DRF parameters: $B = 100$, $\ell = 1000$, $\beta = 0.9$, consider 10 randomly sampled features to split, minimal node size of 5.

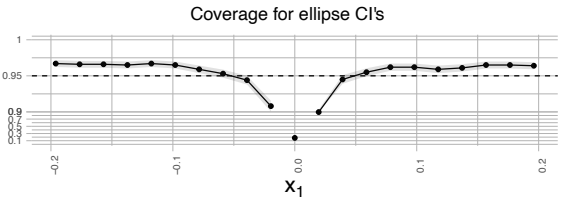


Figure 5.6.4: Ellipsoid confidence intervals for the vector of the three conditional quantiles from Figure 5.6.3. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1000 simulation runs.

CDF on $[0,1]^d$, satisfying

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y} \mid \mathbf{X} = \mathbf{x}) = C_{\mathbf{x}}\big(F_{Y_1 \mid \mathbf{X}=\mathbf{x}}(y), \ldots, F_{Y_d \mid \mathbf{X}=\mathbf{x}}(y)\big).$$

The DRF algorithm may estimate conditional copulas fully nonparametrically or by estimating the parameters of a certain parametric model. For example, if the data comes from a conditional Gaussian copula $\mathbf{Y} \mid \mathbf{X} = \mathbf{x} \sim C^{\text{Gauss}}_{\rho(\mathbf{x})}$, it is enough to estimate the conditional correlation function $\rho(\mathbf{x})$ that characterizes distributional heterogeneity. This is a difficult task because distributional heterogeneity may come from the interdependence of the marginal CDF's due to the copula and may not exclusively occur in the marginals. Because the MMD splitting criterion of DRF is a distributional metric, DRF is able to detect multivariate distributional changes (Gretton et al., 2007).

Subsequently, we consider the conditional Gaussian copula $\mathbf{Y} = (Y_1, Y_2) \mid \mathbf{X} = \mathbf{x} \sim C^{\text{Gauss}}_{\rho(\mathbf{x})}$ with $\mathbf{X} = (X_1, \ldots, X_5) \sim U(-1,1)^5$ and the conditional correlation function $\rho(\mathbf{x}) = \text{Cor}(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x}) = x_1$. That is, both Y_1 and Y_2 follow a standard Gaussian distribution $N(0,1)$ marginally, but their conditional correlation is characterized by $\rho(\mathbf{x}) = x_1$. Ćevid et al. (2022) use a slightly different data generating mechanism because they consider a uniform distribution of the covariates with the support $[0,1]$ instead of $[-1,1]$. We consider $[-1,1]$ such that that the conditional correlation at $x_1 = 0$ does not lie at the boundary of the considered x_1-values because this would artificially introduce boundary effects similar to the conditional quantile estimation above.

We estimate and make inference for $\rho(\mathbf{x}) = x_1$ for a range of values x_1 and different sample sizes n. Figure 5.6.5 illustrates our results. For a sample size of $n = 5000$ (displayed in red), our two-sided DRF confidence intervals achieve the nominal 95% coverage rate for x_1-values that are not too close to either -1 or 1. For x_1-values, and hence conditional correlation values $\text{Cor}(Y_1, Y_2 \mid \mathbf{X} = \mathbf{x})$, that are close to either -1 or 1, we see some degeneration behavior because these values imply the special cases that Y_1 and Y_2 are completely dependent from each other.

5.6.4 | Witness Function for conditional distributional treatment effect

In Section 5.4, we outlined how to test for distributional differences between two treatment groups and how to compute simultaneous confidence bands for the corresponding conditional witness function. To illustrate the performance of DRF in this use case, we revisit the two data generating mechanisms (5.38) and (5.39) that we considered when we analyzed the CATE in Section 5.6.1. In the first case with data from (5.38), there is no treatment effect, and the treatment ($W = 1$) and the control ($W = 0$) groups are equally distributed. In

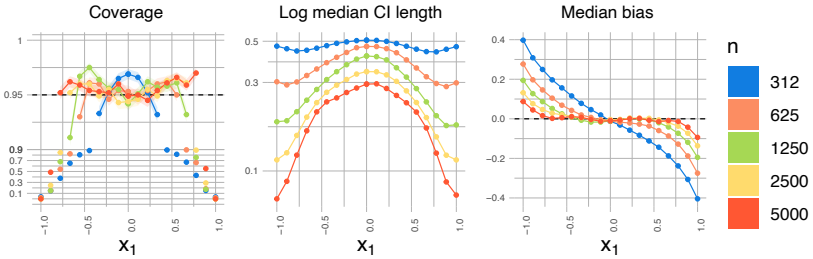


Figure 5.6.5: Estimating conditional correlations $\rho(\mathbf{x}) = \mathrm{Cor}(Y_1, Y_2 \,|\, \mathbf{X} = \mathbf{x})$ of data from the conditional Gaussian copula $\mathbf{Y} = (Y_1, Y_2) \,|\, \mathbf{X} = \mathbf{x} \sim C^{\mathrm{Gauss}}_{\rho(\mathbf{x})}$ with $\mathbf{X} = (X_1, \ldots, X_5) \sim U(-1, 1)^5$ for a range of x_1-values (x-axis) and sample sizes n (differentiated by color). The plot displays the coverage (fraction of times the true, and in general unknown, conditional correlation was inside the confidence interval) and log median length of two-sided 95% confidence intervals for the conditional correlation and median bias over 1000 simulation runs. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1000 simulation runs. In the coverage plot, for $x_1 = -1$ and $x_1 = 1$, the dots from all three values of n are on top of each other. DRF parameters: $B = 100$, $\ell = 1000$, $\beta = 0.9$, consider 10 randomly sampled features to split, minimal node size of 5.

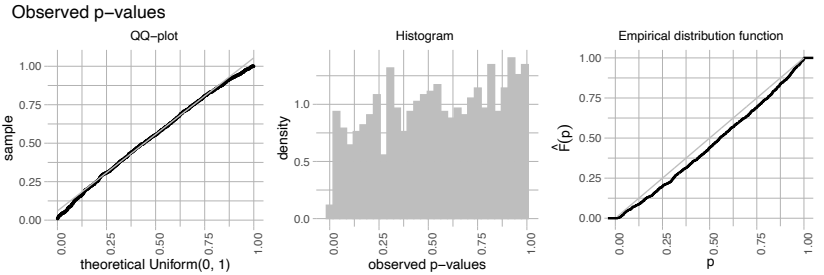
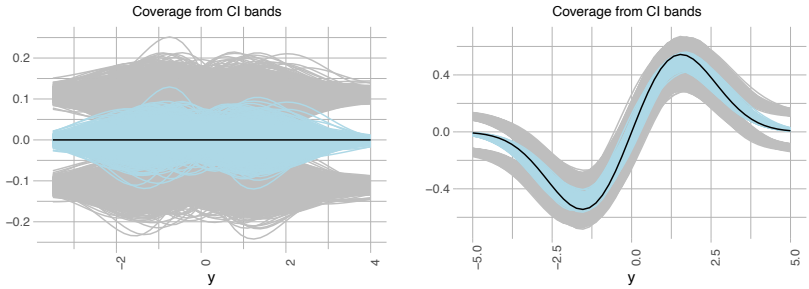


Figure 5.6.6: Two-sided p-values (QQ-plot, histogram, and empirical distribution function) from 1000 repetitions for testing the null hypothesis that the treatment and control groups have equal distributional embeddings at level $\alpha = 5\%$ with data of sample size $n = 5000$ from (5.38) at the test point $\mathbf{x} = (0.7, 0.3, 0.5, 0.68, 0.43)^T$. DRF parameters: $B = 200$, $\ell = 1000$, $\beta = 0.9$, consider 10 randomly sampled features to split, minimal node size of 5.

the second case with data from (5.39), there is a treatment effect.

To formally test if the distributions of the treatment and control groups are different at all, we simulate 1000 data sets of sample size $n = 5000$ each from the two data generating mechanisms and compute the test statistic $\|\hat{\mu}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})\|^2_{\hat{\mathcal{H}}}$ according to (5.36). For each of the 1000 runs, we compute a p-value for testing the null hypothesis that the two embeddings from the treatment and control groups are the same against a two-sided alternative using the approximate bootstrap sample distribution of the test statistic obtained from the B many subforests. Figure 5.6.6 displays our findings for the data generating mechanism (5.38) with equal distributions and illustrates that the p-values are dominated by a Uniform$(0, 1)$ distribution, which is given by the gray line. Consequently, the p-values seem to be valid. In particular, 3.6% (this number has a 95% confidence interval of $(0.0311, 0.0409)$) of them are below the nominal 0.05 level. With the data generating mechanism (5.39), all p-values equal the smallest possible value, and the null hypothesis is always rejected.

To investigate where the treatment and control distributions differ, we estimate the whole conditional witness function and compute simultaneous confidence bands according to (5.37). Figure 5.6.7 illustrates our results. With the data from (5.38) where the treatment and control distributions coincide, 99.8% (95% confidence interval of $(0.9968, 0.9992)$) of the simultaneous 95% confidence bands cover the true underlying conditional witness function that constantly equals 0. Although our method overcovers in this situation, Figure 5.6.7b illustrates that the power goes to 1 under the alternative because no simultaneous confidence band contains the constant zero function. In this



(a) Without treatment effect

(b) With treatment effect

Figure 5.6.7: Simultaneous 95% confidence bands (gray) and conditional witness function estimators (blue) over 1000 repetitions of the true conditional witness function (black) for data of sample size $n = 5000$ without (5.38) in Figure 5.6.7a and with (5.39) in Figure 5.6.7b treatment effect at the test point $\mathbf{x} = (0.7, 0.3, 0.5, 0.68, 0.43)^T$. DRF parameters: $B = 200$, $\ell = 1000$, $\beta = 0.9$, consider 10 randomly sampled features to split, minimal node size of 5.

case, the true conditional witness function is covered in 96.5% of the cases (95% confidence interval of $(0.9601, 0.9699)$).

These simulations illustrate the practical applicability and usefulness of our developments of the conditional distributional treatment effect in Section 5.4. This approach allows us to capture differences between two distributions that may not be represented by mean differences alone. Moreover, our theoretical developments can be directly translated into practice and consequently enable us to perform formal tests that involve test statistics with highly complex and generally intractable distributions.

5.7 | Conclusion

We developed results about the asymptotic distribution of the Distributional Random Forest (DRF) (Ćevid et al., 2022), which is a forest-based (Breiman, 2001) method to nonparametrically estimate Hilbert space embeddings of multivariate conditional distributions in a locally adaptive fashion. The general approach of DRF allows us to estimate a wide range of multivariate targets from one and the same DRF estimator. Because the DRF prediction is Hilbert space-valued, we formulated and developed new theory for Random Forests operating in Hilbert spaces, building on Wager and Athey (2018). In particular, we explicitly characterized the exact asymptotic behavior of the variance of the DRF prediction. Moreover, we established a bootstrap-type result that allowed

us to approximate its distribution in a computationally efficient way.

We presented two strands of applications: we formally tested two treatment groups for distributional differences and investigated where these differences occur, and we estimated and made inference for low-dimensional parameters like the conditional average treatment effect (CATE), conditional quantiles, and conditional correlations. The former application is particularly important to determine differences between the treatment and the control group if the distribution of the two groups are different beyond the mean. To simplify the application of our theory in this former use case, we fitted two DRF's, one for each treatment group, similar to Park et al. (2021). Simulation studies demonstrated the performance and usefulness of our developed inference results for the DRF for these two strands of applications.

Acknowledgements

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

Appendix

5.A | Derivations and Proofs

Preliminaries. First, we recall some of the notation and definitions from the main text. Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote the underlying probability space. Throughout, let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ denote the RKHS associated with the kernel k. We assume that k is bounded and continuous in its two arguments. Boundedness of k ensures that μ is indeed defined on all of $\mathcal{M}_b(\mathbb{R}^d)$, and continuity of $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ ensures that $\mathcal{H}$ is separable. Thus, measurability issues can be avoided. Let us denote by $\xi\colon (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ a map from Ω to $\mathcal{H}$. Separability implies that such a map ξ is measurable if and only if $\langle \xi, f \rangle_{\mathcal{H}}$ is measurable for all $f \in \mathcal{H}$. Moreover, it can easily be checked that $\Phi(P)$ is linear on $\mathcal{M}_b(\mathbb{R}^d)$. Separability of $\mathcal{H}$ and $\mathbb{E}[\|\xi\|_{\mathcal{H}}] < \infty$ mean that the integral

$$\mathbb{E}[\xi] = \int_{\Omega} \xi d\mathbb{P},$$

is well defined and that

$$F(\mathbb{E}[\xi]) = \mathbb{E}[F(\xi)],$$

for any continuous linear function $F\colon \mathcal{H} \to \mathbb{R}$.[1] In particular, $\mathbb{E}[\langle \xi, f \rangle_{\mathcal{H}}] = \langle \mathbb{E}[\xi], f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Moreover, for $q \geq 1$, denote by

$$\begin{aligned}
\mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \{\xi\colon (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H})) \text{ measurable, with } \mathbb{E}[\|\xi\|^q] < \infty]\} \\
\mathbb{L}^q(\Omega, \mathcal{A}, \mathcal{H}) &= \text{Set of equivalence classes in } \mathcal{L}^q(\Omega, \mathcal{A}, \mathcal{H}) \\
\operatorname{Var}(\xi) &= \mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_{\mathcal{H}}^2] = \mathbb{E}[\|\xi\|_{\mathcal{H}}^2] - \|\mathbb{E}[\xi]\|_{\mathcal{H}}^2, \quad \xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}) \\
\operatorname{Cov}(\xi_1, \xi_2) &= \mathbb{E}[\langle \xi_1 - \mathbb{E}[\xi_1], \xi_2 - \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}] = \mathbb{E}[\langle \xi_1, \xi_2 \rangle_{\mathcal{H}}] - \langle \mathbb{E}[\xi_1], \mathbb{E}[\xi_2] \rangle_{\mathcal{H}}, \\
&\qquad \xi_1, \xi_2 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}).
\end{aligned}$$

Furthermore, it is well-known that $(\mathbb{L}^q, \|\cdot\|_{\mathbb{L}^q(\mathcal{H})})$ is a Banach space with

$$\|\xi\|_{\mathbb{L}^q(\mathcal{H})} = \mathbb{E}[\|\xi\|_{\mathcal{H}}^q]^{1/q}.$$

This allows us to also define conditional expectations. For a sub σ-algebra $\mathcal{F} \subset \mathcal{A}$ and an element $\xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H})$, the conditional expectation $\mathbb{E}[\xi \mid \mathcal{F}]$ is the (a.s.) unique element such that
(C1) $\mathbb{E}[\xi \mid \mathcal{F}]\colon (\Omega, \mathcal{F}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable and $\mathbb{E}[\xi \mid \mathcal{F}] \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathcal{H})$,
(C2) $\mathbb{E}[\xi \mathbb{1}_F] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}] \mathbb{1}_F]$ for all $F \in \mathcal{F}$;
see for instance Umegaki and Bharucha-Reid (1970) or Pisier (2016, Chapter

[1] Here and below, $F(\xi)$ is meant to denote $F(\xi(\omega))$ for all $\omega \in \Omega$.

1). Particularly, condition (C2) implies that $\mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]] = \mathbb{E}[\mathbb{E}[\xi \mid \mathcal{F}]\mathbb{1}_\Omega] = \mathbb{E}[\xi]$ due to $\Omega \in \mathcal{F}$ for any σ-algebra. It can also be shown that $F(\mathbb{E}[\xi \mid \mathcal{F}]) = \mathbb{E}[F(\xi) \mid \mathcal{F}]$ for all linear and continuous $F\colon \mathcal{H} \to \mathbb{R}$ and that $\|\mathbb{E}[\xi \mid \mathcal{F}]\|_{\mathcal{H}} \leq \mathbb{E}[\|\xi\|_{\mathcal{H}} \mid \mathcal{F}]$ (Pisier, 2016, Chapter 1). Moreover, it can be shown that

(C3) For $\xi \in \mathbb{L}^2(\Omega, \mathcal{A}, \mathcal{H})$, $\mathbb{E}[\xi \mid \mathcal{F}]$ is the orthogonal projection onto $\mathbb{L}^2(\Omega, \mathcal{F}, \mathcal{H})$;

see (Umegaki and Bharucha-Reid, 1970). Although the conditional expectation $\mathbb{E}[\xi \mid \mathcal{F}]$, similarly to real-valued conditional expectations, is only defined a.s., we do not explicitly state this in our developments below.

We denote by $\mathbb{E}[\xi \mid \mathbf{X}] = \mathbb{E}[\xi \mid \sigma(\mathbf{X})]$. The following Proposition shows that this notion is well defined and establishes further properties of Hilbert space-valued conditional expectations.

Proposition 5.A.1 (Proposition 6 in Ćevid et al. (2022))**.** *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two separable Hilbert spaces, $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$, and $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.*[2]

(C4) There exists a measurable function $h\colon (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ such that $\mathbb{E}[\xi \mid \sigma(\mathbf{X})] = h(\mathbf{X}) = \mathbb{E}[\xi \mid \mathbf{X}]$.

(C5) For $\xi_1 \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi_2 \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_1)$, $\mathbb{E}[\langle\xi_1, \xi_2\rangle_{\mathcal{H}_1} \mid \mathbf{X}] = \langle\mathbb{E}[\xi_1 \mid \mathbf{X}], \xi_2\rangle_{\mathcal{H}_1}$ holds.

(C6) If $\mathbf{X}_2$ and $(\xi, \mathbf{X}_1)$ are independent, then $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$.

(C7) $\mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] \mid \mathbf{X}_1] = \mathbb{E}[\mathbb{E}[\xi \mid \mathbf{X}_1] \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$.

Condition (C4) in particular allows us to consider $\mathbb{E}[\xi \mid \sigma(\mathbf{X})]$ as a function in $\mathbf{X}$ and thus justifies the notation $\mathbb{E}[\xi \mid \mathbf{X}]$ and all the subsequent derivations. We may also define conditional independence through conditional expectation: with the notation of Proposition 5.A.1, ξ and $\mathbf{X}_1$ are conditionally independent given $\mathbf{X}_2$ if $\mathbb{E}[f(\xi) \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[f(\xi) \mid \mathbf{X}_1]$ for all bounded and measurable $f\colon (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$; see Constantinou and Dawid (2017, Proposition 2.3). This leads to two further important properties:

Proposition 5.A.2 (Proposition 7 in Ćevid et al. (2022))**.** *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two separable Hilbert spaces, $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_1)$, and $\xi_1, \xi_2, \xi \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathcal{H}_2)$.*

(C8) If ξ and $\mathbf{X}_2$ are conditionally independent given $\mathbf{X}_1$, then $\mathbb{E}[\xi \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbb{E}[\xi \mid \mathbf{X}_1]$,

(C9) If ξ_1, ξ_2 are conditionally independent given $\mathbf{X}$, then $\mathbb{E}[\langle\xi_1, \xi_2\rangle \mid \mathbf{X}] = \langle\mathbb{E}[\xi_1 \mid \mathbf{X}], \mathbb{E}[\xi_2 \mid \mathbf{X}]\rangle$.

For $\mathbf{x} \in \mathbb{R}^p$, denote by $P_{\mathbf{x}}$ the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ on $\mathbb{R}^d$. For two functions f and g with $\liminf_{s\to\infty} g(s) > 0$, we denote $f(s) = \mathcal{O}(g(s))$ if

$$\limsup_{s\to\infty} \frac{|f(s)|}{g(s)} \leq C$$

[2] We recall that all equalities technically only hold almost surely.

for some $C > 0$. If $C = 1$, we write $f(s) \precsim g(s)$. For a sequence of random variables $X_n \colon \Omega \to \mathbb{R}$ and a sequence of real numbers $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write $X_n = \mathcal{O}_p(a_n)$ if

$$\lim_{M \to \infty} \sup_n \mathbb{P}(a_n^{-1} |X_n| > M) = 0,$$

that is, X_n is bounded in probability. We write $X_n = o_p(a_n)$ if $a_n^{-1} X_n$ converges to zero in probability. Similarly, for (S, d) a separable metric space, $\mathbf{X}_n \colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$, $n \in \mathbb{N}$ and $\mathbf{X} \colon (\Omega, \mathcal{A}) \to (S, \mathcal{B}(S))$ measurable, we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, if $d(\mathbf{X}_n, \mathbf{X}) = o_p(1)$.

Finally, let $\mathbf{X} \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_1)$ and $\xi \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathcal{H}_2)$, and assume that $A \subset \Omega$ depends on $\mathbf{X}$, $A = A(\mathbf{X})$. Thus, for $\mathbf{X}$ fixed to a certain value, A is a fixed set. If $\mathbb{P}(A \mid \mathbf{X}) > 0$ almost everywhere, we define

$$\mathbb{E}[\xi \mid A] = \mathbb{E}[\xi \mid \mathbf{X}, A] = \frac{\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}]}{\mathbb{P}(A \mid \mathbf{X})} \in \mathcal{L}^2(\Omega, \sigma(\mathbf{X}), \mathcal{H}_2).$$

Then, we have by construction that

$$\mathbb{E}[\xi \mathbb{1}_A \mid \mathbf{X}] = \mathbb{E}[\xi \mid \mathbf{X}, A] \cdot \mathbb{P}(A \mid \mathbf{X}). \tag{5.41}$$

Let again $\Phi(\mathbf{x}) = \Phi(P_{\mathbf{x}})$ be the embedding of the true conditional distribution into $\mathcal{H}$. It has the following three properties.

Lemma 5.A.3 (Lemma 8 in Ćevid et al. (2022))**.** *It holds that* $\mathbb{E}[\Phi(\delta_{\mathbf{Y}}) \mid \mathbf{X} = \mathbf{x}] = \Phi(P_{\mathbf{x}})$.

For a more compact notation in the following Lemma, let $N = \{1, \ldots, n\}$, and let for $A \subset N$ and $k \leq |A|$, let $C_k(A)$ be the set of all subsets of size k drawn from A without replacement, with $C_0 = \emptyset$. The following lemma presents a U-statistic expansion that we afterwards apply to an individual tree of our DRF forest.

Lemma 5.A.4 (Lemma 9 in Ćevid et al. (2022))**.** *Let* $(\mathcal{H}_1, \langle \cdot, \cdot \rangle_1)$ *and* $(\mathcal{H}_2, \langle \cdot, \cdot \rangle_2)$ *be two separable Hilbert spaces, and let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be i.i.d. copies of a random element* $\mathbf{Z} \colon (\Omega, \mathcal{A}) \to (\mathcal{H}_1, \mathcal{B}(\mathcal{H}_1))$*. Write* $\mathcal{Z}_n = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$*, and let* $T \colon (\mathcal{H}_1^n, \mathcal{B}(\mathcal{H}_1^n)) \to (\mathcal{H}_2, \mathcal{B}(\mathcal{H}_2))$ *measurable with* $\mathbb{E}[\|T(\mathcal{Z}_n)\|_{\mathcal{H}_2}^2] < \infty$*. If* T *is symmetric, there exist functions* T_j*,* $j = 1, \ldots, n$*, such that*

$$T(\mathcal{Z}_n) = \mathbb{E}[T(\mathbf{Z})] + \sum_{i=1}^n T_1(\mathbf{Z}_i) + \sum_{i_1 < i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \cdots + T_n(\mathcal{Z}_n), \tag{5.42}$$

and it holds that

$$\mathrm{Var}(T(\mathcal{Z}_n)) = \sum_{i=1}^{n} \binom{n}{i} \mathrm{Var}(T_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)) \tag{5.43}$$

and

$$T_1(\mathbf{Z}_i) = \mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)].$$

Subsequently, we apply this expansion to an individual tree of our DRF predictor. Let $\hat{\mu}_n(\mathbf{x})$ be as in (5.5), namely

$$\hat{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < i_2 < \ldots < i_{s_n}} \mathbb{E}_\varepsilon \left[T(\mathbf{x}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}) \right], \tag{5.44}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsamples $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}$ of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and $s_n \to \infty$ with n and where

$$T(\mathbf{x}, \varepsilon; \mathbf{Z}_1, \ldots, \mathbf{Z}_{s_n}) = \sum_{j=1}^{s_n} \frac{\mathbb{1}(\mathbf{X}_j \in \mathcal{L}(\mathbf{x}))}{|\mathcal{L}(\mathbf{x})|} \Phi(\delta_{\mathbf{Y}_j}).$$

We introduce the following additional notation similar to Section 5.3. Let $\mathcal{Z}_{s_n} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{s_n})$ concatenate s_n i.i.d. copies of $\mathbf{Z}$, and define for $j = 1, \ldots, s_n$

$$\begin{aligned} \mathrm{Var}(T) &= \mathrm{Var}(T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})), \\ \mathrm{Var}(T_j) &= \mathrm{Var}(\mathbb{E}[T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \ldots, \mathbf{Z}_j]) \end{aligned}$$

We note that, due to i.i.d. sampling, what kind of subset $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}$ we are considering affects neither variance nor expectation. In particular, we might always take $\mathcal{Z}_{s_n}$. Using composition (5.42) on $\hat{\mu}_n(\mathbf{x})$ gives

$$\begin{aligned} \hat{\mu}_n(\mathbf{x}) = \mathbb{E}[T(\mathcal{Z}_{s_n})] + \binom{n}{s_n}^{-1} \Bigg(\binom{n-1}{s_n - 1} \sum_{i=1}^{n} T_1(\mathbf{Z}_i) + \binom{n-2}{s_n - 2} \sum_{i_1 < i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) \\ + \ldots + \sum_{i_1 < i_2 < \ldots < i_{s_n}} T_{s_n}(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}) \Bigg). \end{aligned} \tag{5.45}$$

This representation was used in Ćevid et al. (2022) to prove that the variance of $\hat{\mu}_n(\mathbf{x})$ can be bounded by the scaled variance of a single tree:

Lemma 5.A.5 (Lemma 10 in Ćevid et al. (2022))**.** *Let $\hat{\mu}_n(\mathbf{x})$ be as in* (5.44)*, and assume $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ satisfies **(F3)** and $\mathrm{Var}(T) < \infty$. Then,*

$$\mathrm{Var}(\hat{\mu}_n(\mathbf{x})) \leq \frac{s_n^2}{n} \mathrm{Var}(T_1) + \frac{s_n^2}{n^2} \mathrm{Var}(T) \tag{5.46}$$

$$\leq \left(\frac{s_n}{n} + \frac{s_n^2}{n^2}\right) \mathrm{Var}(T). \tag{5.47}$$

Subsequently, we derive a first-order approximation of the whole forest and of an individual tree. In the following, we denote the second element of (5.45) by

$$\tilde{\mu}_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \binom{n-1}{s_n-1} \sum_{i=1}^{n} T_1(\mathbf{Z}_i) = \frac{s_n}{n} \sum_{i=1}^{n} T_1(\mathbf{Z}_i), \tag{5.48}$$

which is the first order approximation of $\mu_n(\mathbf{x})$ Similarly, applying (5.45) to a tree $T(\mathcal{Z}_{s_n}) = \mathbb{E}_\varepsilon\left[T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})\right]$, we obtain the expansion

$$T(\mathcal{Z}_{s_n}) = \mathbb{E}[T(\mathcal{Z}_{s_n})] + \sum_{i=1}^{s_n} T_1(\mathbf{Z}_i) + \sum_{i_1<i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \ldots T_{s_n}(\mathcal{Z}_n).$$

Consequently, we define

$$\tilde{T}(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} T_1(\mathbf{Z}_i) = \sum_{i=1}^{s_n} \mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{X}_i] - \mathbb{E}[T(\mathcal{Z}_n)]. \tag{5.49}$$

Contrary to $T(\mathcal{Z}_{s_n})$, $\tilde{T}(\mathcal{Z}_{s_n})$ is a sum of independent random elements on $\mathcal{H}$ and thus much easier to handle. A key argument will thus be to show that $\tilde{T}(\mathcal{Z}_{s_n})$ approximates $T(\mathcal{Z}_{s_n})$ asymptotically.

Consider the leaf $\mathcal{L}(\mathbf{x})$ of the tree $T(\mathcal{Z}_{s_n})$ that contains the test point $\mathbf{x}$. To emphasize the dependence of such a leaf node on the training data, we will sometimes write $\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})$ instead of $\mathcal{L}(\mathbf{x})$ in the following.

As in Meinshausen (2006); Wager and Athey (2017), the crucial part of proving that a Random Forest is consistent is to establish that the diameter of the leaf $\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})$ goes to zero in probability. In particular, we need a refined result from Wager and Athey (2017):

Lemma 5.A.6 (Lemma 2 of Wager and Athey (2018))**.** *Let T be a tree satisfying **(F2)** and **(F4)** that is trained on data $\mathcal{Z}_{s_n} = (\xi_1, \mathbf{X}_1), \ldots, (\xi_{s_n}, \mathbf{X}_{s_n})$, and let $\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})$ be the leaf of $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ containing $\mathbf{x}$. Suppose that assumption **(D1)** holds for $\mathbf{X}_1, \ldots, \mathbf{X}_{s_n}$. Then,*

$$\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \geq \sqrt{p}\left(\frac{s_n}{2k-1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) \leq p\left(\frac{s_n}{2k-1}\right)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}. \tag{5.50}$$

Lemma 5.A.7 (Lemma 12 in Ćevid et al. (2022))**.** *Let $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ be a tree satisfying **(F1)** and **(F5)**, and let $\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})$ be the leaf of $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$*

containing $\mathbf{x}$*. Then,*

$$\mathbb{E}[T(\mathcal{Z}_{s_n})] = \mathbb{E}[\mathbb{E}[\xi_1 \mid \mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})]] \tag{5.51}$$

and

$$\mathrm{Var}(T(\mathcal{Z}_{s_n})) \leq \sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]. \tag{5.52}$$

Corollary 5.A.8 (Corollary 13 in Ćevid et al. (2022))**.** *In addition to the conditions of Lemma 5.A.6, assume* ***(D2)*** *and that the trees* $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ *in the forest satisfy* ***(F1)*** *and* ***(F4)****. Then, we have*

$$\|\mathbb{E}[\hat{\mu}_n(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left(s_n^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) \tag{5.53}$$

and

$$\|\mathbb{E}[\xi \mid \mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})]\|_{\mathcal{H}} \xrightarrow{p} \|\mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}. \tag{5.54}$$

If moreover ***(D3)*** *holds, then we have*

$$\mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})] \xrightarrow{p} \mathbb{E}[\|\xi\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]. \tag{5.55}$$

Lemma 5.A.9. *Let* $\xi_{1,n}, \xi_{2,n} \in \mathcal{L}^2(\Omega, \mathcal{A}, H)$ *for* $n \in \mathbb{N}$*, and assume that we have*

(I) $\mathrm{Var}(\xi_{1,n}) = \mathcal{O}(g_1(n))$ *and* $\mathrm{Var}(\xi_{2,n}) = \mathcal{O}(g_1(n))$*,*

(II) $\mathrm{Var}(\xi_{1,n} - \xi_{2,n}) = \mathcal{O}(g_2(n))$

for some functions $g_1, g_2 \colon \mathbb{N} \to \mathbb{N}$*. Then,* $|\mathrm{Var}(\xi_{1,n}) - \mathrm{Var}(\xi_{2,n})| = \mathcal{O}(g_2(n)) + \mathcal{O}(\sqrt{g_1(n)}\sqrt{g_2(n)})$*.*

Proof. It holds that

$$\begin{aligned}\left|\sqrt{\mathrm{Var}(\xi_{1,n})} - \sqrt{\mathrm{Var}(\xi_{2,n})}\right| &= |\|\xi_{1,n} - \mathbb{E}[\xi_{1,n}]\|_{\mathcal{L}^2} - \|\xi_{2,n} - \mathbb{E}[\xi_{2,n}]\|_{\mathcal{L}^2}| \\ &\leq \|\xi_{1,n} - \xi_{2,n} - (\mathbb{E}[\xi_{1,n}] - \mathbb{E}[\xi_{2,n}])\|_{\mathcal{L}^2} \\ &= \sqrt{\mathrm{Var}(\xi_{1,n} - \xi_{2,n})}, \end{aligned} \tag{5.56}$$

where we used the reverse triangle inequality in the second step. Thus, we in particular have $\sqrt{\mathrm{Var}(\xi_{1,n})} \leq \sqrt{\mathrm{Var}(\xi_{2,n})} + \sqrt{\mathrm{Var}(\xi_{1,n} - \xi_{2,n})}$ or

$$\mathrm{Var}(\xi_{1,n}) \leq \mathrm{Var}(\xi_{2,n}) + \mathrm{Var}(\xi_{1,n} - \xi_{2,n}) + 2\sqrt{\mathrm{Var}(\xi_{2,n})}\sqrt{\mathrm{Var}(\xi_{1,n} - \xi_{2,n})}.$$

Symmetrically, it holds that

$$\mathrm{Var}(\xi_{2,n}) \leq \mathrm{Var}(\xi_{1,n}) + \mathrm{Var}(\xi_{1,n} - \xi_{2,n}) + 2\sqrt{\mathrm{Var}(\xi_{1,n})}\sqrt{\mathrm{Var}(\xi_{1,n} - \xi_{2,n})}$$

so that by assumption $\text{Var}(\xi_{1,n}) - \text{Var}(\xi_{2,n}) = \mathcal{O}(g_2(n)) + \mathcal{O}(\sqrt{g_1(n)}\sqrt{g_2(n)})$. □

Define in the following the number of data points belonging to the same leaf as $\mathbf{x}$ as $N_{\mathbf{x}} = |\{j : \mathbf{X}_j \in \mathcal{L}(\mathbf{x})\}|$ and let

$$S_i = \frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x})\}}{N_{\mathbf{x}}}, \tag{5.57}$$

be the weight associated with each observation i in a tree $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$, such that

$$T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} S_i k(\mathbf{Y}_i, \cdot).$$

We will make use the following property of the S_i:

$$1 = \mathbb{E}\left[\sum_{i=1}^{s_n} S_i\right] = \sum_{i=1}^{s_n} \mathbb{E}[S_i] = s_n \mathbb{E}[S_1]. \tag{5.58}$$

In particular,

$$\text{Var}(\mathbb{E}[S_1|\mathbf{X}_1]) \leq \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1]^2] \leq \mathbb{E}[\mathbb{E}[S_1|\mathbf{X}_1]] = \mathbb{E}[S_1] = \mathcal{O}(s_n^{-1}) \tag{5.59}$$

Lemma 5.A.10 (Lemma 4 of Wager and Athey (2017) slightly adapted)**.** *Suppose $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed on $[0,1]^p$ with a density f that is bounded away from infinity, and let $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ be α-regular **(F4)**. Then, there is a constant $C_{f,p}$ depending on f and p such that,*

$$s_n \text{Var}(\mathbb{E}[S_1|\mathbf{Z}_1]) \gtrsim \frac{1}{\kappa} \frac{C_{f,p}}{\log(s_n)} \tag{5.60}$$

When f is uniform over $[0,1]^p$, the bound holds with $C_{f,p} = 2^{-(p+1)}(p-1)!$

Let $\tilde{T}(\mathcal{Z}_{s_n})$ be the first order approximation of $T(\mathcal{Z}_{s_n}) = \mathbb{E}_\varepsilon[T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})]$ as in (5.49). We now prove that the variance of $\tilde{T}(\mathcal{Z}_{s_n})$ does not decrease to zero too fast compared to the variance of $T(\mathcal{Z}_{s_n})$, which is a key result that allows us to meaningfully approximate $T(\mathcal{Z}_{s_n})$ with $\tilde{T}(\mathcal{Z}_{s_n})$. The main result in (5.62) is called $\nu(s_n)$-incrementality of the tree $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ in Wager and Athey (2017, Definition 6). Before we introduce the result, we note that, due to the orthogonal decomposition in (5.45), we have

$$\text{Var}(\tilde{T}(\mathcal{Z}_{s_n})) = s_n \text{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \leq \text{Var}(T(\mathcal{Z}_{s_n})).$$

Thus in particular, if $\text{Var}(T(\mathcal{Z}_{s_n})) < \infty$, we also have $\text{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathcal{O}(s_n^{-1})$.

Theorem 5.A.11. *Suppose that the tree $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ satisfies **(F1)** and **(F4)**. Suppose in addition that **(D1) – (D4)** hold. Then,*

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \gtrsim \mathrm{Var}(\mathbb{E}[S_1|\mathbf{Z}_1])\mathrm{Var}(\xi|\mathbf{X}=\mathbf{x}) \tag{5.61}$$

and

$$\frac{\mathrm{Var}(\tilde{T}(\mathcal{Z}_{s_n}))}{\mathrm{Var}(T(\mathcal{Z}_{s_n}))} \gtrsim \frac{C_{f,p}}{\log(s_n)^p}, \tag{5.62}$$

where $C_{f,p}$ is the constant from Lemma 5.A.10.

Proof. Consider the concatenated data $\mathcal{Z}_{s_n} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{s_n})$. First, assume (5.61) is true. In this case, we know from Lemma 5.A.10 that

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1]) \gtrsim \frac{1}{\kappa}\frac{\nu(s_n)}{s_n}\mathrm{Var}(\xi \mid \mathbf{X}=\mathbf{x}),$$

where $\nu(s) = \frac{C_{f,p}}{\log(s)}$. By Corollary 5.A.8, it holds that $\mathbb{E}[\|\xi\|_{\mathcal{H}}^2|\mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})] \xrightarrow{p} \mathbb{E}[\|\xi\|_{\mathcal{H}}^2|\mathbf{X}=\mathbf{x}]$, so that

$$\begin{aligned}\mathrm{Var}(\xi|\mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) &= \mathbb{E}[\|\xi\|_{\mathcal{H}}^2|\mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})] - \|\mathbb{E}[\xi|\mathbf{X} \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})]\|_{\mathcal{H}}^2 \\ &\xrightarrow{p} \mathrm{Var}(\xi \mid \mathbf{X}=\mathbf{x}).\end{aligned}$$

Thus, using the same argument as in the proof of Theorem 5 in Wager and Athey (2017), $\mathrm{Var}(T(\mathcal{Z}_{s_n})) \precsim \mathrm{Var}(\xi \mid \mathbf{X}=\mathbf{x})/k$. Consequently, due to i.i.d. sampling, we have

$$\frac{\mathrm{Var}(\tilde{T}(\mathcal{Z}_{s_n}))}{\mathrm{Var}(T(\mathcal{Z}_{s_n}))} = \frac{s_n\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1])}{\mathrm{Var}(T(\mathcal{Z}_{s_n}))} \gtrsim \nu(s),$$

which establishes the result.

Before we verify (5.61), we note that, as we use double-sampling, separate data is used for prediction ($\mathcal{I}$) and leaf building ($\mathcal{I}^c$). Consequently, $\mathbf{Z}_1$ might fall into the prediction set, $1 \in \mathcal{I}$, or the leave building set, $1 \notin \mathcal{I}$. However, only the former case may contribute to the variance:

Claim: For some $\varepsilon > 0$,

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1]) = \mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, 1 \in \mathcal{I}]) + \mathcal{O}(s_n^{-(1+\varepsilon)}) \tag{5.63}$$

Proof: By assumption, we have $\mathbb{P}(1 \in \mathcal{I} \mid \mathbf{Z}_1) = \mathbb{P}(1 \in \mathcal{I}) = 1/2$ for each tree.

Thus, we have

$$\begin{aligned}\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1] &= \mathbb{E}[T(\mathcal{Z}_{s_n})\mathbb{1}\{1 \in \mathcal{I}\} \mid \mathbf{Z}_1] + \mathbb{E}[T(\mathcal{Z}_{s_n})\mathbb{1}\{1 \notin \mathcal{I}\} \mid \mathbf{Z}_1] \\ &= \frac{1}{2}\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}] + \frac{1}{2}\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}],\end{aligned}$$

and consequently

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1]) \\ =&\frac{1}{4}\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}]) + \frac{1}{4}\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}]) \\ &+ \frac{1}{2}\mathrm{Cov}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}], \mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}]).\end{aligned}$$

Next, using analogous arguments as in Wager and Athey (2017, Corollary 6), we have

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}]) = \mathcal{O}\left(s_n^{-(1+C_\alpha \frac{\pi}{p})}\right)$$

with $C_\alpha = \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$. Finally, since from above

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}]) \leq \mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) = \mathcal{O}(s_n^{-1}),$$

it follows that

$$\begin{aligned}&|\mathrm{Cov}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}], \mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}])| \\ \leq&(\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \in \mathcal{I}\}])\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, \{1 \notin \mathcal{I}\}]))^{1/2} \\ =&\mathcal{O}(s_n^{-(1+1/2C_\alpha \frac{\pi}{p})}).\end{aligned}$$

Choosing $\varepsilon = 1/2C_\alpha \frac{\pi}{p} > 0$ gives the result. □

Because the tree T satisfies **(F1)** and **(F4)** and due to assumption **(D1)**, we can apply Lemma 5.A.10. Thus, once (5.61) is proven, Lemma 5.A.10 and (5.63) imply

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1, 1 \in \mathcal{I}]) \gtrsim C\frac{1}{s_n \log(s_n)},$$

so that the remainder term in (5.63) is negligible. Consequently, we assume for the remainder of the proof that $1 \in \mathcal{I}$ and absorb the randomness due to the data $\{\mathbf{Z}_i \colon i \notin \mathcal{I}\}$ for building the leaves into the randomness of the tree. In addition, we also write s_n instead of $s_n/2$ in the tree predictions. That is, we write $T(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} S_i'\xi_i$ although $T(\mathcal{Z}_{s_n}) = \sum_{i\in\mathcal{I}} S_i'\xi_i$ with $|\mathcal{I}| = s_n/2$ is technically correct. With (5.63) and i.i.d. sampling, this simply amounts to a

change of constants.

In the remainder of the proof, we verify (5.61). Note that due to honesty, we have

$$\mathrm{Var}(\mathbb{E}[S_1 \mid \mathbf{Z}_1]) = \mathrm{Var}(\mathbb{E}[S_1 \mid \xi_1, \mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1 \mid \mathbf{X}_1]). \tag{5.64}$$

Thus, it is enough to prove (5.61) with $\mathrm{Var}(\mathbb{E}[S_1 \mid \mathbf{X}_1])$. To do this, we use a truncation trick from Wager and Athey (2017). We define

$$T'(\mathcal{Z}_{s_n}) = T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \leq s_n^{-w}\}, \tag{5.65}$$

$$S_i' = S_i \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \leq s_n^{-w}\}, \text{ where } w = \frac{1}{2}\frac{\pi}{p}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}, \tag{5.66}$$

so that $T'(\mathcal{Z}_{s_n}) = \sum_{i=1}^{s_n} S_i' \xi_i$. Crucially, w is chosen such that

$$\mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}) = \mathcal{O}(s_n^{-w}) \tag{5.67}$$

This follows from Lemma 5.A.6, as in Wager and Athey (2017).

Claim: (5.61) holds for T'.

Proof:

We start first with a variance lower bound:

Claim:

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1]) \\
=&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1]) + \mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \xi_1, \mathbf{X}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1]) \\
\geq&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \xi_1, \mathbf{X}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1]).
\end{aligned} \tag{5.68}$$

Proof: We need to prove the first equality and start with the decomposition

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{Z}_1]) \\
=&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \xi_1, \mathbf{X}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] + \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1]).
\end{aligned}$$

Consider for $\mathcal{A} = \sigma(\sigma(\mathbf{X}_1), \sigma(\xi_1))$ the space

$$\mathbb{L}^2(\Omega, \sigma(\mathbf{X}_1), H) \subset \mathbb{L}^2(\Omega, \mathcal{A}, H).$$

This space is a Hilbert space with the inner product

$$\langle \xi_1, \xi_2 \rangle_{\mathbb{L}^2} = \mathbb{E}[\langle \xi_1, \xi_2 \rangle_H].$$

Moreover, $\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] = \mathbb{E}[\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathcal{A}] \mid \mathbf{X}_1]$ is a projection from $\mathbb{E}[T'(\mathcal{Z}_{s_n})] \mid \mathcal{A}] \in \mathbb{L}^2(\Omega, \mathcal{A}, H)$ to $\mathbb{L}^2(\Omega, \sigma(\mathbf{X}_1), H)$. Thus, we have

$$\begin{aligned}&\mathrm{Cov}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1], \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \xi_1, \mathbf{X}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1])\\ =&\langle \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1], \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \xi_1, \mathbf{X}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] \rangle_{\mathbb{L}^2} = 0.\end{aligned}$$

□

Now, by honesty (i) ξ_i is independent of S_i' conditional on $\mathbf{X}_i$, and more generally, (ii) ξ_i is independent of S_j', $j = 1, \ldots, n$, conditional on $\mathbf{X}_i$. Thus, using (i), (ii), and the independence of ξ_1 from ξ_j, $j > 1$, we have

$$\begin{aligned}\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, \xi_1] &= \mathbb{E}[S_1'\xi_1 \mid \mathbf{X}_1, \xi_1] + \sum_{i=2}^{n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1, \xi_1]\\ &= \mathbb{E}[S_1' \mid \mathbf{X}_1, \xi_1]\mathbb{E}[\xi_1 \mid \mathbf{X}_1, \xi_1] + \sum_{i=2}^{n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1]\\ &= \mathbb{E}[S_1' \mid \mathbf{X}_1]\xi_1 + \sum_{i=2}^{n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1]\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] &= \mathbb{E}[S_1'\xi_1 \mid \mathbf{X}_1] + \sum_{i=2}^{n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1]\\ &= \mathbb{E}[S_1' \mid \mathbf{X}_1]\mathbb{E}[\xi_1 \mid \mathbf{X}_1] + \sum_{i=2}^{n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1]\end{aligned}$$

and consequently

$$\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1, \xi_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mathbb{E}[\xi_1 \mid \mathbf{X}_1])). \tag{5.69}$$

Furthermore, we can refine this statement to:
Claim:

$$\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mathbb{E}[\xi_1 \mid \mathbf{X}_1])) = \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))) + \mathcal{O}(s_n^{-(1+2w)}), \tag{5.70}$$

where w is defined as in (5.66).
Proof:
We have

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mathbb{E}[\xi_1 \mid \mathbf{X}_1]))\\ =&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mathbb{E}[\xi_1 \mid \mathbf{X}_1] + \mu(\mathbf{x}) - \mu(\mathbf{x})))\\ =&\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x})) - \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x})))\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))) + \mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x})))\\
&\quad - \mathrm{Cov}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))). \qquad (5.71)
\end{aligned}
$$

Because $\mathbb{E}[S_1' \mid \mathbf{X}_1]$ maps into $\mathbb{R}_{\geq 0}$, we have

$$
\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))) &\leq \mathbb{E}[\|\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1]) - \mu(\mathbf{x}))\|_{\mathcal{H}}^2]\\
&= \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1]^2 \|\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x})\|_{\mathcal{H}}^2]\\
&\leq \mathbb{E}[\mathbb{E}[S_1'^2 \mid \mathbf{X}_1] C^2 \|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p}^2]\\
&\leq \mathbb{E}[S_1'^2] C^2 s_n^{-2w}, \qquad (5.72)
\end{aligned}
$$

where we used assumption **(D2)** for the third inequality and where the last step followed because $\mathbb{E}[S_1'^2 \mid \mathbf{X}_1] = 0$, for $\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of $S_1' = S_1 \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \leq s_n^{-w}\}$. Due to similar arguments, we have

$$
\begin{aligned}
&|\mathrm{Cov}(\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x})))|\\
=&|\mathbb{E}[\langle \mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))\rangle]\\
&\quad - \langle \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))], \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))]\rangle|\\
\leq&|\mathbb{E}[\langle \mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x})), \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))\rangle]|\\
&\quad + |\langle \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))], \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))]\rangle|\\
=&|\mathbb{E}[\langle \mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x})), \mathbb{E}[S_1' | \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))\rangle]|\\
&\quad + |\langle \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))], \mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))]\rangle|\\
=&\mathbb{E}[\|\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))\|_{\mathcal{H}}^2] + \|\mathbb{E}[\mathbb{E}[S_1' \mid \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))]\|_{\mathcal{H}}^2\\
\leq&\mathbb{E}[S_1'^2] C^2 s_n^{-2w} + \mathbb{E}[\|\mathbb{E}[S_1' | \mathbf{X}_1](\mathbb{E}[\xi_1 \mid \mathbf{X}_1] - \mu(\mathbf{x}))\|_{\mathcal{H}}^2]\\
\leq&2\mathbb{E}[S_1'^2] C^2 s_n^{-2w} \qquad (5.73)
\end{aligned}
$$

Observe that we have

$$
\mathbb{E}[S_1'^2] \leq \mathbb{E}[S_1^2] \leq \mathbb{E}[S_1] = \frac{1}{s_n} \sum_{i=1}^{s_n} \mathbb{E}[S_i] = \frac{1}{s_n} \mathbb{E}\left[\sum_{i=1}^{s_n} S_i\right] = \frac{1}{s_n},
$$

due to $S_1 \in [0, 1]$, $\mathbb{E}[S_1] = \ldots = \mathbb{E}[S_n]$ and $\sum_{i=1}^{s} S_i = 1$. Finally, combining this observation with (5.72) and (5.73) gives (5.70).

□

Next, we establish

Claim:

$$
\begin{aligned}
&\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))) \qquad (5.74)\\
=&\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \mathrm{Var}(\xi_1 | \mathbf{X} = \mathbf{x}) + \mathcal{O}(s_n^{-(1+w)}) + \mathcal{O}(s_n^{-2}). \qquad (5.75)
\end{aligned}
$$

Proof:

We have

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))) \\
=&\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]^2\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}^2] - \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}]^2 \\
=&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]^2\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X}_1]\right] - \mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}|\mathbf{X}_1]\right]^2.
\end{aligned} \tag{5.76}$$

The second term in (5.76) can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}|\mathbf{X}_1])\right]^2 \\
\leq&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\|\xi_1\|_{\mathcal{H}} + \|\mu(\mathbf{x})\|_{\mathcal{H}}|\mathbf{X}_1]\right]^2 \\
=&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\left(\mathbb{E}[\|\xi_1\|_{\mathcal{H}}|\mathbf{X}_1] + \|\mu(\mathbf{x})\|_{\mathcal{H}}\right)\right]^2 \\
\leq&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\left(\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi_1\|_{\mathcal{H}}|\mathbf{X} = \mathbf{x}] + \|\mu(\mathbf{x})\|_{\mathcal{H}}\right)\right]^2 \\
=&\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]\right]^2\left(\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi_1\|_{\mathcal{H}}|\mathbf{X} = \mathbf{x}] + \|\mu(\mathbf{x})\|_{\mathcal{H}}\right)^2 \\
=&\mathbb{E}[S_1']^2\left(\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi_1\|_{\mathcal{H}}|\mathbf{X} = \mathbf{x}] + \|\mu(\mathbf{x})\|_{\mathcal{H}}\right)^2 \\
=&\mathcal{O}(s_n^{-2}).
\end{aligned}$$

The last step followed because of (5.10), a consequence of **(D1)** and **(D3)**. The first term in (5.76) can be bounded by

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}[S_1'|\mathbf{X}_1]^2\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X}_1]\Big] \\
=&\mathbb{E}\Big[\mathbb{E}[S_1'|\mathbf{X}_1]^2\big(\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X}_1] - \mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X} = \mathbf{x}] \\
&\quad + \mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X} = \mathbf{x}]\big)\Big] \\
=&\mathbb{E}\Big[\mathbb{E}[S_1'|\mathbf{X}_1]^2\left(\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X}_1] - \mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|^2|\mathbf{X} = \mathbf{x}]\right)\Big] \\
&\quad + \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]^2]\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x}).
\end{aligned} \tag{5.77}$$

Because S_1' is defined as $S_1' = S_1\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \leq s_n^{-w}\}$, it is zero if $\|\mathbf{X}_1 - \mathbf{x}\|_{\mathbb{R}^p} > s_n^{-w}$. Combining this with Assumption **(D2)** and **(D3)** it follows that

$$\left|\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]^2(\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}^2|\mathbf{X}_1] - \mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}])\right]\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]^2 \left|\mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathbf{X}_1] - \mathbb{E}[\|\xi_1 - \mu(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]\right|\right] \\
&= \mathbb{E}\big[\mathbb{E}[S_1'|\mathbf{X}_1]^2 |\mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X}_1] + \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \\
&\qquad - 2\langle \mathbb{E}[\xi \mid \mathbf{X}_1], \mu(\mathbf{x})\rangle_{\mathcal{H}} - \mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}] \\
&\qquad - \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 + 2\langle \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}], \mu(\mathbf{x})\rangle_{\mathcal{H}}\big] \\
&\leq \mathbb{E}\big[\mathbb{E}[S_1'|\mathbf{X}_1]^2 \big(\big|\mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X}_1] - \mathbb{E}[\|\xi_1\|_{\mathcal{H}}^2 \mid \mathbf{X} = \mathbf{x}]\big| \\
&\qquad + 2\big|\langle \mathbb{E}[\xi \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[\xi \mid \mathbf{X}_1], \mu(\mathbf{x})\rangle_{\mathcal{H}}\big|\big)\big] \\
&\leq \mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]^2\right] (C_1 s_n^{-w} + C_2 s_n^{-w}) \\
&= \mathcal{O}(s_n^{-(1+w)})
\end{aligned} \tag{5.78}
$$

holds, where we used $\mathbb{E}\left[\mathbb{E}[S_1'|\mathbf{X}_1]^2\right] = \mathcal{O}(s_n^{-1})$. Finally, due to

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]^2] =& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) + \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]]^2 \\
=& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) + \mathbb{E}[S_1']^2 \\
=& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) + \mathcal{O}(1/s_n^2),
\end{aligned} \tag{5.79}
$$

we can combine (5.76)-(5.79) to establish our claim (5.74).

□

Combining (5.68), (5.69), (5.70), and (5.74), we get that (5.61) holds for T' due to

$$
\begin{aligned}
&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \\
\geq& \mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1, \xi_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1]) \\
=& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\xi_1 - \mathbb{E}[\xi_1|\mathbf{X}_1])) \\
=& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\xi_1 - \mu(\mathbf{x}))) + \mathcal{O}(s_n^{-(1+2w)}) \\
=& \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x}) + \mathcal{O}(s_n^{-(1+w)}) + \mathcal{O}(s_n^{-2}) + \mathcal{O}(s_n^{-(1+2w)}).
\end{aligned} \tag{5.80}
$$

□

In the next step we replace S_1' with S_1 in the expression above.
Claim:

$$
|\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) - \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])| = \mathcal{O}(s_n^{-(1+w/2)}). \tag{5.81}
$$

Proof: We have

$$
\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1] - \mathbb{E}[S_1'|\mathbf{X}_1]) &= \mathrm{Var}(\mathbb{E}[S_1 - S_1'|\mathbf{X}_1]) \\
&= \mathrm{Var}(\mathbb{E}[S_1 \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{X}_1]) \\
&\leq \mathbb{E}[\mathbb{E}[S_1 \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{X}_1]^2]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}[S_1 \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}] \\
&= \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}[S_i \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}] \\
&= \frac{1}{s} \mathbb{E}[\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\} \sum_{i=1}^{s} S_i] \\
&= \frac{1}{s} \mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}) \\
&= \mathcal{O}(s_n^{-(1+w)}) \qquad (5.82)
\end{aligned}
$$

due to $\sum_{i=1}^{s_n} S_i = 1$ and where the last step followed due to (5.67). As $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1]) = \mathcal{O}(s_n^{-1})$ from (5.59) and analogously $\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) = \mathcal{O}(s_n^{-1})$, it holds by Lemma 5.A.9 and (5.82) that

$$
|\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) - \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])| = \mathcal{O}(s_n^{-(1+w)}) + \mathcal{O}(s_n^{-((2+w)/2)}) = \mathcal{O}(s_n^{-(1+w/2)}).
$$

□

Thus, we have

$$
\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \geq \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x}) + \mathcal{O}(s_n^{-(1+\varepsilon)}), \quad (5.83)
$$

for some $\varepsilon > 0$. Because we have $\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x}) > 0$ by assumption and due to $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1|\mathbf{Z}_1]) \gtrsim C(s_n \log(s_n))^{-1}$ by Lemma 5.A.10, we finally have

$$
\liminf_{n\to\infty} \frac{\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1])}{\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x})} \geq 1, \qquad (5.84)
$$

or (5.61) for $T'(\mathcal{Z}_{s_n})$ instead of $T(\mathcal{Z}_{s_n})$.

Now, it also holds that:

Claim:

$$
|\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) - \mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1])| = \mathcal{O}(s_n^{-(1+w/2)}) \qquad (5.85)
$$

Proof: First, observe that we have

$$
\begin{aligned}
&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1] - \mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \\
=&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{Z}_1]).
\end{aligned}
$$

Using composition (5.42) on $T''(\mathcal{Z}_{s_n}) = T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}$, we have

$$
T''(\mathcal{Z}_{s_n}) = \mathbb{E}[T''(\mathcal{Z}_{s_n})] + \sum_{i=1}^{s_n} T_1''(\mathbf{Z}_i) + \sum_{i_1<i_2} T_2''(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \cdots + T_{s_n}''(\mathcal{Z}_{s_n}),
$$

$$\begin{aligned}\mathrm{Var}(T''(\mathcal{Z}_{s_n})) &= \sum_{i=1}^{s_n} \binom{s_n}{i} \mathrm{Var}(T_i''(\mathbf{Z}_1, \dots, \mathbf{Z}_i)),\\ T_1''(\mathbf{Z}_1) &= \mathbb{E}[T''(\mathcal{Z}_{s_n})|\mathbf{Z}_1] - \mathbb{E}[T''(\mathcal{Z}_{s_n})],\end{aligned}$$

and thus

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{Z}_1])\\ =&\mathrm{Var}(T_1''(\mathbf{Z}_1))\\ \leq&\frac{1}{s_n}\sum_{i=1}^{s_n}\binom{s_n}{i}\mathrm{Var}(T_i''(\mathbf{Z}_1, \dots, \mathbf{Z}_i))\\ =&\frac{1}{s_n}\mathrm{Var}(T''(\mathcal{Z}_{s_n}))\\ =&\frac{1}{s_n}\mathrm{Var}(T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}).\end{aligned}$$

Moreover, with analogous arguments as in the proof of Lemma 12 in Ćevid et al. (2022) it can be shown that,

$$\begin{aligned}&\mathrm{Var}(T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\})\\ \leq&\mathbb{E}\left[\left\|\sum_{i=1}^{s} S_i\xi_i\right\|_{\mathcal{H}}^2 \mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}\right]\\ \leq&C\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi\|_{\mathcal{H}}^2|\mathbf{X}=\mathbf{x}]\mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w})\end{aligned}$$

such that

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w}\}|\mathbf{Z}_1])\\ \leq&\frac{C\sup_{\mathbf{x}\in[0,1]^p}\mathbb{E}[\|\xi\|_{\mathcal{H}}^2|\mathbf{X}=\mathbf{x}]}{s_n}\mathbb{P}(\mathrm{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) > s_n^{-w})\\ =&\mathcal{O}(s_n^{-(1+w)}),\end{aligned}$$

where the last step follows from (5.67) and (5.10). As also $\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \leq \mathrm{Var}(T(\mathcal{Z}_{s_n}))/s_n = \mathcal{O}(s_n^{-1})$ and similarly for T', the claim holds by Lemma 5.A.9, similar to the proof of (5.81) above. □

Summarizing everything, it follows from (5.85), (5.80) and (5.81),

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1])\\ =&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) + \mathcal{O}(s_n^{-(1+w/2)})\\ \geq&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X}=\mathbf{x}) + \mathcal{O}(s_n^{-(1+w)}) + \mathcal{O}(s_n^{-2}) + \mathcal{O}(s_n^{-(1+2w)})\\ =&\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X}=\mathbf{x}) + \mathcal{O}(s_n^{-(1+w)}) + \mathcal{O}(s_n^{-2}) + \mathcal{O}(s_n^{-(1+2w)}).\end{aligned}$$

Because $\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1]) \gtrsim C(s_n \log(s_n))^{-1}$ by Lemma 5.A.10 and $\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x}) > 0$ by assumption, this implies that

$$\liminf_{s_n} \frac{\mathrm{Var}(\mathbb{E}[T(\mathbf{Z})|\mathbf{Z}_1])}{\mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x})} \geq 1, \tag{5.86}$$

or $\mathrm{Var}(\mathbb{E}[T(\mathbf{Z})|\mathbf{Z}_1]) \gtrsim \mathrm{Var}(\mathbb{E}[S_1|\mathbf{X}_1])\mathrm{Var}(\xi_1|\mathbf{X} = \mathbf{x})$, proving (5.61).

□

Theorem 5.3.2. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Denote by $\mathbf{Z}_i = (\mathbf{X}_i, k(\mathbf{Y}_i, \cdot))$, $i = 1, \ldots, n$. Then, there exists a map $T_n \colon [0,1]^p \times \mathcal{H} \to \mathcal{H}$ such that, with*

$$\sigma_n^2 = \frac{s_n^2}{n}\mathrm{Var}(T_n(\mathbf{Z}_1)), \tag{5.11}$$

we have $\sigma_n \to 0$, $\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\| = \mathcal{O}_p(\sigma_n)$, and

$$\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) = \frac{s_n}{n}\sum_{i=1}^{n} T_n(\mathbf{Z}_i) + o_p(\sigma_n). \tag{5.12}$$

Moreover, T_n is given by

$$T_n(\mathbf{Z}_i) = \mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)]. \tag{5.13}$$

Proof. Let $\tilde{\mu}_n(\mathbf{x})$ and $\tilde{T}(\mathcal{Z}_{s_n})$ be as in (5.48) and (5.49), respectively, and observe that we have

$$\sigma_n^2 = \mathrm{Var}(\tilde{\mu}(\mathbf{x})) = \frac{s_n^2}{n}\mathrm{Var}(T_1) = \frac{s_n}{n}s_n\mathrm{Var}(T_1) = \frac{s_n}{n}\mathrm{Var}(\tilde{T}(\mathcal{Z}_{s_n})) \leq \frac{s_n}{n}\mathrm{Var}(T).$$

We first prove (5.12) for $\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]$.
Claim: (5.12) holds for $\mathbb{E}[\hat{\mu}_n(\mathbf{x})]$ in place of $\mu_n(\mathbf{x})$:

$$\hat{\mu}(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})] = \frac{s_n}{n}\sum_{i=1}^{n}(\mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)]) + o_p(\sigma_n) \tag{5.87}$$

Proof: First,
Claim:

$$\frac{1}{\sigma_n^2}\mathbb{E}[\|\hat{\mu}_n(\mathbf{x}) - \tilde{\mu}_n(\mathbf{x})\|_{\mathcal{H}}^2] \precsim \frac{s_n}{n}\frac{\log(s_n)^p}{C_{f,p}} \to 0 \tag{5.88}$$

Proof: Let $(s_n)_j = s_n(s_n - 1)\cdots(s_n - (j-1)) = s_n!/(s_n - j)!$ and $\mathrm{Var}(\tilde{T}) =$

$\mathrm{Var}(\tilde{T}(\mathcal{Z}_{s_n}))$. Then, using the decomposition in (5.45) with

$$\mathrm{Var}(T_j) = \mathrm{Var}(T_j(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_j)), \quad j = 1, \ldots, s_n$$

that

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\mathbb{E}[\|\hat{\mu}(\mathbf{x}) - \tilde{\mu}(\mathbf{x})\|_{\mathcal{H}}^2] \\
=&\frac{1}{\sigma_n^2}\mathrm{Var}\bigg(\binom{n}{s_n}^{-1}\bigg(\binom{n-2}{s_n-2}\sum_{i_1<i_2} T_2(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}) + \cdots \\
&\quad + \sum_{i_1<i_2<\ldots<i_{s_n}} T_{s_n}(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}})\bigg)\bigg) \\
=&\frac{1}{\sigma_n^2}\sum_{i=2}^{s_n}\left(\frac{(s_n)_i}{(n)_i}\right)^2\binom{n}{i}\mathrm{Var}(T_i) \\
=&\frac{1}{\sigma_n^2}\sum_{i=2}^{s_n}\left(\frac{(s_n)_i}{(n)_i}\right)\binom{s_n}{i}\mathrm{Var}(T_i) \\
\leq&\frac{1}{\sigma_n^2}\frac{(s_n)_2}{(n)_2}\sum_{i=2}^{s_n}\binom{s_n}{i}\mathrm{Var}(T_i) \\
\leq&\frac{s_n^2}{n^2}\frac{\mathrm{Var}(T)}{\sigma_n^2} \\
=&\frac{s_n}{n}\frac{\mathrm{Var}(T)}{\mathrm{Var}(\tilde{T})} \\
\lesssim&\frac{s_n}{n}\frac{\log(s_n)^p}{C_{f,p}},
\end{aligned}$$

where we used Theorem 5.A.11 in the last step. Finally, since $s_n = n^\beta$ for $\beta < 1$, we infer $(s_n \log(s_n)^p)/n \to 0$. □

Since by construction $\mathbb{E}[\hat{\mu}_n(\mathbf{x})] = \mathbb{E}[\tilde{\mu}_n(\mathbf{x})]$, for all $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{\sigma_n}(\hat{\mu}(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})]) - \frac{1}{\sigma_n}(\tilde{\mu}(\mathbf{x}) - \mathbb{E}[\tilde{\mu}(\mathbf{x})])\right\|_{\mathcal{H}}^2 > \varepsilon\right) \\
\leq&\frac{1}{\varepsilon^2}\frac{1}{\sigma_n^2}\mathbb{E}[\|\hat{\mu}(\mathbf{x}) - \tilde{\mu}(\mathbf{x})\|_{\mathcal{H}}^2].
\end{aligned}$$

Consequently, we have $\|\frac{1}{\sigma_n}(\hat{\mu}(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})]) - \frac{1}{\sigma_n}(\tilde{\mu}(\mathbf{x}) - \mathbb{E}[\tilde{\mu}(\mathbf{x})])\|_{\mathcal{H}} \to 0$ in probability, or equivalently

$$\hat{\mu}(\mathbf{x}) - \mathbb{E}[\hat{\mu}(\mathbf{x})] = \tilde{\mu}(\mathbf{x}) - \mathbb{E}[\tilde{\mu}(\mathbf{x})] + o_p(\sigma_n).$$

Since moreover

$$\tilde{\mu}(\mathbf{x}) - \mathbb{E}[\tilde{\mu}(\mathbf{x})] = \frac{s_n}{n}\sum_{i=1}^{n} T_1(\mathbf{Z}_i) = \frac{s_n}{n}\sum_{i=1}^{n}(\mathbb{E}[T(\mathcal{Z}_n) \mid \mathbf{Z}_i] - \mathbb{E}[T(\mathcal{Z}_n)]),$$

we conclude Claim (5.87). □

Due to

$$\frac{1}{\sigma_n}\left\|\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right\|_{\mathcal{H}} \leq \frac{1}{\sigma_n}\left\|\hat{\mu}_n(\mathbf{x}) - \mathbb{E}[\tilde{\mu}(\mathbf{x})]\right\|_{\mathcal{H}} + \frac{1}{\sigma_n}\left\|\mathbb{E}[\tilde{\mu}(\mathbf{x})] - \mu(\mathbf{x})\right\|_{\mathcal{H}},$$

the result follows if we can show that the second expression in this upper bound goes to zero

$$\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \gtrsim \frac{1}{\kappa}\frac{C_{f,p}}{s_n \log(s_n)^p}\mathrm{Var}(\xi|\mathbf{X}=\mathbf{x}) > 0,$$

so that

$$\begin{aligned}
\sigma_n^2 =& \frac{s_n^2}{n}\mathrm{Var}(T_1) \\
=& \frac{s_n^2}{n}\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]) \\
\gtrsim& \frac{s_n^2}{n}\frac{1}{\kappa}\frac{C_{f,p}}{s_n \log(s_n)^p}\mathrm{Var}(\xi|\mathbf{X}=\mathbf{x}) \\
=& \frac{s_n}{n\log(s_n)^p}\mathrm{Var}(\xi|\mathbf{X}=\mathbf{x})\frac{1}{\kappa}C_{f,p}.
\end{aligned}$$

Thus, using that $s_n = n^\beta$, we have

$$\sigma_n = \Omega\left(\frac{\sqrt{s_n}}{\sqrt{n\log(s_n)^p}}\right) = \Omega\left(\left(\frac{n^\beta}{n\beta^p\log(n)^p}\right)^{1/2}\right) = \Omega\left(\left(n^{\beta-1-\varepsilon}\right)^{1/2}\right)$$

for some $\varepsilon > 0$ On the other hand, due to Theorem 5.A.8, we have

$$\|\mathbb{E}[\hat{\mu}(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}} = \mathcal{O}\left(s_n^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) = \mathcal{O}\left(s_n^{-1/2C_\alpha\frac{\pi}{p}}\right) = \mathcal{O}\left(n^{-1/2\beta C_\alpha\frac{\pi}{p}}\right),$$

which implies

$$\frac{\|\mathbb{E}[\hat{\mu}(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}}}{\sigma_n} = \mathcal{O}\left(n^{-1/2(\beta C_\alpha\frac{\pi}{p}+\beta-1-\varepsilon)}\right) = \mathcal{O}\left(n^{-1/2(\beta(1+C_\alpha\frac{\pi}{p})-1-\varepsilon)}\right).$$

This goes to zero provided that $-(\beta(1 + C_\alpha\frac{\pi}{p}) - 1 - \varepsilon) < 0$ or $\beta >$

$(1+\varepsilon)\left(1+C_\alpha\frac{\pi}{p}\right)^{-1}$, which is satisfied for $\varepsilon>0$ small enough if

$$\beta > \left(1+C_\alpha\frac{\pi}{p}\right)^{-1}.$$

Taking $T_n(\mathbf{Z}_i)=\mathbb{E}[T(\mathcal{Z}_n)\mid\mathbf{Z}_i]-\mathbb{E}[T(\mathcal{Z}_n)]$ gives the claimed result. □

Before being able to prove Theorem 5.3.5 in the main text, we need to refine the characterization of the asymptotic behavior of the variance of $T_n(\mathbf{Z}_i)$.

Theorem 5.3.4. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then, for all $f\in\mathcal{H}\setminus\{0\}$, we have*

$$\lim_{n\to\infty}\frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1),f\rangle)}{\mathrm{Var}(T_n(\mathbf{Z}_1))}=\frac{\mathrm{Var}(\langle k(\mathbf{Y},\cdot),f\rangle|\mathbf{X}=\mathbf{x})}{\mathrm{Var}(k(\mathbf{Y},\cdot)|\mathbf{X}=\mathbf{x})}=\sigma^2(f)>0. \tag{5.15}$$

Proof. Note that, due to (5.63), we can again "ignore" the double-sampling and assume to condition on a point $\mathbf{Z}_1$ with index in the prediction set $\mathcal{I}$ and use s_n instead of $s_n/2$ elements in the tree predictions. First, due to $T_n(\mathbf{Z}_1)=\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1]-\mathbb{E}[T(\mathcal{Z}_{s_n})]$, we infer

$$\frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1),f\rangle)}{\mathrm{Var}(T_n(\mathbf{Z}_1))}=\frac{\mathrm{Var}(\mathbb{E}[\langle T(\mathcal{Z}_{s_n}),f\rangle\mid\mathbf{Z}_1])}{\mathrm{Var}(\mathbb{E}[T(\mathcal{Z}_{s_n})|\mathbf{Z}_1])}.$$

Combining (5.68) with (5.80) in Theorem 5.A.11, we have

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}),f\rangle\mid\mathbf{Z}_1])\\
=&\mathrm{Var}(\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}),f\rangle\mid\mathbf{X}_1])+\mathrm{Var}(\mathbb{E}[S_1'\mid\mathbf{X}_1])\mathrm{Var}(\langle\xi_1,f\rangle\mid\mathbf{X}=\mathbf{x})+\mathcal{O}(s_n^{-(1+\epsilon)}),\\
&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{Z}_1])\\
=&\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1])+\mathrm{Var}(\mathbb{E}[S_1'\mid\mathbf{X}_1])\mathrm{Var}(\xi_1\mid\mathbf{X}=\mathbf{x})+\mathcal{O}(s_n^{-(1+\epsilon)})
\end{aligned} \tag{5.89}$$

for some $\epsilon>0$. Let in the following $\mathbb{1}_{w,s_n}=\mathbb{1}\{\mathrm{diam}(\mathcal{L}(\mathbf{x},\mathcal{Z}_{s_n}))\le s_n^{-w}\}$ such that $S_i'=S_i\mathbb{1}_{w,s_n}$. We now show that
Claim:

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}),f\rangle\mid\mathbf{X}_1])&=\mathcal{O}(s_n^{-(1+\epsilon)})\\
\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n})|\mathbf{X}_1])&=\mathcal{O}(s_n^{-(1+\epsilon)}).
\end{aligned} \tag{5.90}$$

Proof: First, due to honesty, we have

$$\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}),f\rangle\mid\mathbf{X}_1]=\mathbb{E}[S_1'\mid\mathbf{X}_1]\mathbb{E}[\langle\xi_1,f\rangle\mid\mathbf{X}_1]+\sum_{i=2}^{s_n}\mathbb{E}[S_i'\langle\xi_i,f\rangle\mid\mathbf{X}_1]$$

and

$$\mathbb{E}[T'(\mathcal{Z}_{s_n}) \mid \mathbf{X}_1] = \mathbb{E}[S_1' \mid \mathbf{X}_1]\mathbb{E}[\xi_1 \mid \mathbf{X}_1] + \sum_{i=2}^{s_n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1].$$

Subsequently, we consider the variance of the two terms and their covariance individually. First, we study the variance of the first terms. The variances satisfy

Claim:

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\xi_1|\mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[\xi_1|\mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + \mathcal{O}(s_n^{-(1+w)}) \tag{5.91}$$

and

$$\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\langle \xi_1, f\rangle|\mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle \xi_1, f\rangle|\mathbf{X} = \mathbf{x}]^2 + \mathcal{O}(s_n^{-(1+w)}). \tag{5.92}$$

Proof:

We only show (5.91) because (5.92) follows analogously. We have

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\xi_1|\mathbf{X}_1]) \\
=&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}) + \mu(\mathbf{x}))) \\
=&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) + \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))) \\
&+ \mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))\right).
\end{aligned} \tag{5.93}$$

Because

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) =&\mathbb{E}[\|(\mathbb{E}[S_1'|\mathbf{X}_1] - \mathbb{E}[S_1'])\mu(\mathbf{x})\|_{\mathcal{H}}^2] \\
=&\mathbb{E}[(\mathbb{E}[S_1'|\mathbf{X}_1] - \mathbb{E}[S_1'])^2]\|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \\
=&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2,
\end{aligned}$$

it follows that

$$\begin{aligned}
&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\xi_1|\mathbf{X}_1]) \\
=&\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2 + \mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))) \\
&+ \mathrm{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))\right).
\end{aligned}$$

Because $\mathbb{E}[S_1'|\mathbf{X}_1]$ maps into $\mathbb{R}_{\geq 0}$, we have

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))) \leq& \mathbb{E}[\|\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1]) - \mu(\mathbf{x}))\|_{\mathcal{H}}^2] \\
=& \mathbb{E}[\mathbb{E}[S_1'|\mathbf{X}_1]^2\|(\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))\|_{\mathcal{H}}^2]
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}[\mathbb{E}[S_1'^2|\mathbf{X}_1]C^2\|\mathbf{X}_1 - x\|_{\mathbb{R}^p}^2] \\
&\leq \mathbb{E}[S_1'^2]C^2 s_n^{-2w},
\end{aligned}$$

where the last step followed because $\mathbb{E}[S_1'^2|\mathbf{X}_1] = 0$, for $\|\mathbf{X}_1 - x\|_{\mathbb{R}^p} > s_n^{-w}$ by definition of $S_1' = S_1 \mathbb{1}\{\text{diam}(\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \leq s_n^{-w}\}$. Since $\mathbb{E}[S_1'^2] \leq \mathbb{E}[S_1'] \leq \mathbb{E}[S_1] = \mathcal{O}(s_n^{-1})$ from (5.58), we have

$$\text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))) = \mathcal{O}(s_n^{-(1+2w)}).$$

Finally, we infer

$$\begin{aligned}
&|\text{Cov}\left(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}), \mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x}))\right)| \\
\leq &\sqrt{\text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x}))}\sqrt{\text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1](\mathbb{E}[\xi_1|\mathbf{X}_1] - \mu(\mathbf{x})))} \\
= &\mathcal{O}(s_n^{-(1+w)}),
\end{aligned}$$

due to

$$\text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})) = \text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1]) \cdot \mathcal{O}(1) = \mathcal{O}(s_n^{-1}), \tag{5.94}$$

again using (5.59). Thus, our Claim (5.91) holds. □

Before we continue proving the theorem, we note that, due to honesty, we have

$$\begin{aligned}
\sum_{i=2}^{s_n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1] &= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i'\xi_i \mid \mathbf{X}_i, \mathbf{X}_1] \mid \mathbf{X}_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i' \mid \mathbf{X}_i, \mathbf{X}_1]\mathbb{E}[\xi_i \mid \mathbf{X}_i, \mathbf{X}_1] \mid \mathbf{X}_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[\mathbb{E}[S_i'\mathbb{E}[\xi_i \mid \mathbf{X}_i] \mid \mathbf{X}_i, \mathbf{X}_1] \mid \mathbf{X}_1] \\
&= \sum_{i=2}^{s_n} \mathbb{E}[S_i'\mathbb{E}[\xi_i \mid \mathbf{X}_i] \mid \mathbf{X}_1].
\end{aligned} \tag{5.95}$$

Now, we consider the variance of the sum in (5.95):

Claim:

$$\text{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1]\right) = \text{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mathbb{E}[\xi_1|\mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + \mathcal{O}(s_n^{-(1+w)}) \tag{5.96}$$

and

$$\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\langle\xi_i,f\rangle\mid\mathbf{X}_1]\right)=\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle\xi_1,f\rangle|\mathbf{X}=\mathbf{x}]^2+\mathcal{O}(s_n^{-(1+w)}). \tag{5.97}$$

Proof:

First we note that, using the definition of S_i', it holds that

$$\begin{aligned}\sum_{i=1}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]&=\mathbb{E}[\sum_{i=1}^{s_n}S_i'\mid\mathbf{X}_1]\\&=\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x},\mathcal{Z}_{s_n}))\leq s_n^{-w}\mid\mathbf{X}_1\right)\\&=\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x},\mathcal{Z}_{s_n}))\leq s_n^{-w}\right),\end{aligned}$$

where the last step follows from **(F1)** and the fact that $1\in\mathcal{I}$. Thus abbreviating $p_n=\mathbb{P}\left(\mathrm{diam}(\mathcal{L}(\mathbf{x},\mathcal{Z}_{s_n}))\leq s_n^{-w}\right)$, it follows that

$$\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]=p_n-\mathbb{E}[S_1'\mid\mathbf{X}_1]] \tag{5.98}$$

We only show (5.96), because (5.97) follows analogously. By (5.98) and since p_n is a constant,

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2&=\mathrm{Var}((p_n-\mathbb{E}[S_1'|\mathbf{X}_1])\mu(\mathbf{x}))\\&=\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right).\end{aligned}$$

Thus, we need to show that

$$\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\xi_i\mid\mathbf{X}_1]\right)=\mathrm{Var}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)+\mathcal{O}(s_n^{-(1+w)}), \tag{5.99}$$

which according to Lemma 5.A.9 is implied by

$$\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\xi_i\mid\mathbf{X}_1]-\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1]\right)=\mathcal{O}(s_n^{-(1+2w)}), \tag{5.100}$$

$$\mathrm{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\xi_i\mid\mathbf{X}_1]\right)=\mathcal{O}(s_n^{-1}), \tag{5.101}$$

and (5.94). Subsequently, we establish (5.100) and (5.101). Now, with (5.95),

we have

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\xi_i \mid \mathbf{X}_1] - \mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i' \mid \mathbf{X}_1]\right)\\
=&\operatorname{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mathbb{E}[\xi_i \mid \mathbf{X}_i] \mid \mathbf{X}_1] - \mathbb{E}[S_i'\mu(\mathbf{x}) \mid \mathbf{X}_1]\right)\\
=&\operatorname{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'(\mathbb{E}[\xi_i \mid \mathbf{X}_i] - \mu(\mathbf{x})) \mid \mathbf{X}_1]\right)\\
=&\operatorname{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]\right), \qquad (5.102)
\end{aligned}$$

with $\Delta(\mathbf{X}_i) = \mathbb{E}[\xi_i \mid \mathbf{X}_i] - \mu(\mathbf{x})$. Next, we note that for each i, we have

$$\begin{aligned}
&\mathbb{E}[S_i'\Delta(\mathbf{X}_i) \mid \mathbf{X}_1]\\
=&\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1] + \mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mathbb{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1].
\end{aligned}$$

With $N_j = j + \sum_{i=2}^{s_n}\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}$, $j \in \{0, 1\}$, we have

$$\begin{aligned}
&\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\\
=&\mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\mathbb{1}_{w,s_n}}{N_1}\Delta(\mathbf{X}_i)\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\,\middle|\, \mathbf{X}_1\right]\\
=&\mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\mathbb{1}_{w,s_n}}{N_1}\Delta(\mathbf{X}_i)\,\middle|\, \mathbf{X}_1, \{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\right]\\
&\cdot \mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1)\\
=&\mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\mathbb{1}_{w,s_n}}{N_1}\Delta(\mathbf{X}_i)\right]\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1),
\end{aligned}$$

where the last step follows due to independence of $\mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})$ and $\mathbf{X}_1$ by **(F1)**. Define the element

$$E_i^1 = \mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}\mathbb{1}_{w,s_n}}{N_1}\Delta(\mathbf{X}_i)\right].$$

Because this is nonrandom element of $\mathcal{H}$ and $\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1)$ does not depend on the index $i \in \mathcal{I}$, it follows that

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\right)\\
=&\left\|\sum_{i=2}^{s_n}E_i^1\right\|_{\mathcal{H}}^2 \operatorname{Var}(\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1))
\end{aligned}$$

$$\leq \left(\sum_{i=2}^{s_n} \|E_i^1\|_{\mathcal{H}} \right)^2 \mathbb{E}[\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]^2] \tag{5.103}$$

Due to Jensen's inequality,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]^2] &\leq \mathbb{E}[\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]] \\ &= \mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})) \\ &= \mathcal{O}(s_n^{-1}),\end{aligned} \tag{5.104}$$

where the last step followed because $2\kappa - 1 \geq \mathbb{E}[N_{\mathbf{x}}] = \sum_{i=1}^{s_n} \mathbb{E}[\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}] = s_n \mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}))$ by **(F4)**. On the other hand, we have

$$\begin{aligned}\sum_{i=2}^{s_n} \|E_i^1\|_{\mathcal{H}} &\leq \sum_{i=2}^{s_n} \mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mathbb{1}_{w,s_n}}{N_1} \|\Delta(\mathbf{X}_i)\|_{\mathcal{H}} \right] \\ &\leq \sum_{i=2}^{s_n} \mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mathbb{1}_{w,s_n}}{N_1} C \|\mathbf{X}_i - \mathbf{x}\|_{\mathbb{R}^p} \right] \\ &\leq C s_n^{-w} \mathbb{E}\left[\sum_{i=2}^{s_n} \frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mathbb{1}_{w,s_n}}{N_1} \right] \\ &\leq C s_n^{-w}\end{aligned} \tag{5.105}$$

as $0 \leq \sum_{i=2}^{s_n} \mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}/N_1 \leq 1$. Combining Equations (5.104) and (5.105) with (5.103) gives

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1] \right) = \mathcal{O}(s_n^{-(1+2w)}). \tag{5.106}$$

Similarly, we have

$$\begin{aligned}&\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1] \right) \\ &= \left\| \sum_{i=2}^{s_n} E_i^0 \right\|_{\mathcal{H}}^2 \mathrm{Var}(\mathbb{P}(\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1))\end{aligned} \tag{5.107}$$

with

$$E_i^0 = \mathbb{E}\left[\frac{\mathbb{1}\{\mathbf{X}_i \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mathbb{1}_{w,s_n}}{N_0} \Delta(\mathbf{X}_i) \right] \in \mathcal{H}.$$

With the same arguments as before, it follows that

$$\left\| \sum_{i=2}^{s_n} E_i^0 \right\|_{\mathcal{H}}^2 = \mathcal{O}(s_n^{-2w}).$$

Combining this with

$$\begin{aligned}\mathrm{Var}(\mathbb{P}(\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1)) &= \mathrm{Var}(1 - \mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1)) \\ &= \mathrm{Var}(\mathbb{P}(\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n}) \mid \mathbf{X}_1)) \\ &\leq \mathbb{E}[\mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]^2] \\ &\leq \mathbb{E}[\mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\}] \\ &= \mathcal{O}(s_n^{-1})\end{aligned}$$

results in

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\right) = \mathcal{O}(s_n^{-(1+2w)}). \tag{5.108}$$

Consequently, we have

$$\begin{aligned}&\left|\mathrm{Cov}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1],\right.\right. \\ &\qquad \left.\left.\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\right)\right| \\ \leq &\left(\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \in \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\right)\right. \\ &\qquad \left.\cdot \mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mathbb{1}\{\mathbf{X}_1 \notin \mathcal{L}(\mathbf{x}, \mathcal{Z}_{s_n})\} \mid \mathbf{X}_1]\right)\right)^{1/2} \\ = &\mathcal{O}(s_n^{-(1+2w)}),\end{aligned}$$

so that (5.100) holds. Finally, using the reverse triangle inequality as in (5.56) in the proof of Lemma 5.A.9, we obtain

$$\mathrm{Var}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \xi_i \mid \mathbf{X}_1]\right)^{1/2} = \mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1] \mu(\mathbf{x}))^{1/2} + \mathcal{O}(s_n^{-(1/2+w)}) = \mathcal{O}(s_n^{-1/2}),$$

by (5.100) and (5.94). This shows (5.101) and thus (5.96) in the claim holds true. □

Finally, we consider the covariance between $\mathbb{E}[S_1' \xi_i \mid \mathbf{X}_1]$ and $\sum_{i=2}^{s_n} \mathbb{E}[S_i' \xi_i \mid \mathbf{X}_1]$.

Claim: For some $\varepsilon > 0$, we have

$$\begin{aligned}&\mathrm{Cov}\left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \xi_i \mid \mathbf{X}_1], \mathbb{E}[S_1' | \mathbf{X}_1] \mathbb{E}[\xi_1 | \mathbf{X}_1]\right) \\ = &-\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \|\mathbb{E}[\xi_1 | \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 + \mathcal{O}(s_n^{-(1+\varepsilon)})\end{aligned} \tag{5.109}$$

and

$$\begin{aligned}&\operatorname{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\langle\xi_i,f\rangle\mid\mathbf{X}_1],\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\langle\xi_1,f\rangle|\mathbf{X}_1]\right)\\&=-\operatorname{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\mathbb{E}[\langle\xi_1,f\rangle|\mathbf{X}=\mathbf{x}]^2+\mathcal{O}(s_n^{-(1+\varepsilon)}).\end{aligned}\tag{5.110}$$

Proof: Again, we only show (5.109), because (5.110) follows analogously. Using (5.95), we can subtract and add $\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'|\mathbf{X}_1]$ and $\mathbb{E}[S_1'|\mathbf{X}_1]\mu(\mathbf{x})$ to obtain

$$\begin{aligned}&\operatorname{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\xi_i\mid\mathbf{X}_1],\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\xi_1|\mathbf{X}_1]\right)\\=&\operatorname{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mathbb{E}[\xi_i\mid\mathbf{X}_i]\mid\mathbf{X}_1],\mathbb{E}[S_1'|\mathbf{X}_1]\mathbb{E}[\xi_1|\mathbf{X}_1]\right)\\=&\operatorname{Cov}\Bigg(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1]+\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\\&\qquad\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]+\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\Bigg)\\=&\operatorname{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1],\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]\right)\\&\quad+\operatorname{Cov}\left(\sum_{i=2}^{s_n}\mathbb{E}[S_i'\Delta(\mathbf{X}_i)\mid\mathbf{X}_1],\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\&\quad+\operatorname{Cov}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mathbb{E}[S_1'\Delta(\mathbf{X}_1)|\mathbf{X}_1]\right)\\&\quad+\operatorname{Cov}\left(\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'\mid\mathbf{X}_1],\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)\\=:&(I)+(II)+(III)+(IV),\end{aligned}$$

where again $\Delta(\mathbf{X}_i)=\mathbb{E}[\xi_i\mid\mathbf{X}_i]-\mu(\mathbf{x})$. Since from (5.98),

$$\mu(\mathbf{x})\sum_{i=2}^{s_n}\mathbb{E}[S_i'|\mathbf{X}_1]=\mu(\mathbf{x})(p_n-\mathbb{E}[S_1'|\mathbf{X}_1]),\tag{5.111}$$

it holds that

$$(IV)=\operatorname{Cov}\left(\mu(\mathbf{x})(p_n-\mathbb{E}[S_1'|\mathbf{X}_1]),\mu(\mathbf{x})\mathbb{E}[S_1'|\mathbf{X}_1]\right)=-\operatorname{Var}(\mathbb{E}[S_1'|\mathbf{X}_1])\|\mu(\mathbf{x})\|_{\mathcal{H}}^2.$$

Subsequently, we show that the remaining terms are negligible. Due to the

Cauchy–Schwarz inequality, we have

$$|(I)| \leq \left(\mathrm{Var} \left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mid \mathbf{X}_1] \right) \mathrm{Var}(\mathbb{E}[S_1' \Delta(\mathbf{X}_1) | \mathbf{X}_1]) \right)^{1/2}.$$

As proven above (combining (5.100) and (5.102)), $\mathrm{Var}(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mid \mathbf{X}_1]) = \mathcal{O}(s_n^{-(1+2w)})$, and it can be established that

$$\mathrm{Var}(\mathbb{E}[S_1' \Delta(\mathbf{X}_1) | \mathbf{X}_1]) \leq \mathbb{E}[\mathbb{E}[S_1' \| \Delta(\mathbf{X}_1) \|_{\mathcal{H}} | \mathbf{X}_1]^2] = \mathcal{O}(s_n^{-(1+2w)})$$

holds. Consequently, $(I) = \mathcal{O}(s_n^{-(1+2w)})$. Similarly,

$$|(II)| \leq \left(\mathrm{Var} \left(\sum_{i=2}^{s_n} \mathbb{E}[S_i' \Delta(\mathbf{X}_i) \mid \mathbf{X}_1] \right) \mathrm{Var}(\mu(\mathbf{x}) \mathbb{E}[S_1' | \mathbf{X}_1]) \right)^{1/2} = \mathcal{O}(s_n^{-(1+w)}),$$

as $\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \leq \mathbb{E}[(S_1')^2] = \mathcal{O}(s_n^{-1})$. Finally,

$$\begin{aligned} |(III)| &= |\mathrm{Cov}\left(\mu(\mathbf{x})(1 - \mathbb{E}[S_1' \Delta(\mathbf{X}_1) | \mathbf{X}_1]), \mathbb{E}[S_1' \Delta(\mathbf{X}_1) | \mathbf{X}_1]\right)| \\ &= | - \|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \mathrm{Var}(\mathbb{E}[S_1' \Delta(\mathbf{X}_1) | \mathbf{X}_1])| \\ &= \mathcal{O}(s_n^{-(1+2w)}) \end{aligned}$$

as above.

□

Combining (5.91), (5.96), and (5.109), we obtain

$$\begin{aligned} &\mathrm{Var}(\mathbb{E}[T'(\mathcal{Z}_{s_n}) | \mathbf{X}_1]) \\ =&2\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \|\mathbb{E}[\xi_1 | \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 - 2\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \|\mathbb{E}[\xi_1 | \mathbf{X} = \mathbf{x}]\|_{\mathcal{H}}^2 \\ &+ \mathcal{O}(s_n^{-(1+\epsilon)}) \\ =&\mathcal{O}(s_n^{-(1+\epsilon)}) \end{aligned}$$

and analogously

$$\mathrm{Var}(\mathbb{E}[\langle T'(\mathcal{Z}_{s_n}), f \rangle \mid \mathbf{Z}_1]) = \mathcal{O}(s_n^{-(1+\epsilon)}),$$

proving (5.90).

□

Recall $\mathrm{Var}(\mathbb{E}[S_1' | \mathbf{X}_1]) \sim \mathrm{Var}(\mathbb{E}[S_1 | \mathbf{X}_1]) = \mathrm{Var}(\mathbb{E}[S_1 | \mathbf{Z}_1]) = \Omega((s_n \log(s_n))^{-1})$, by (5.81), (5.64), and Lemma 5.A.10, respectively. This together with Claim (5.90) and the expansion in (5.89) establishes (5.15). □

This leads us to the proof of Theorem 5.3.5 in the main text.

Theorem 5.3.5. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and*

*(**K2**) hold. Then,*

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})\right) \xrightarrow{D} N(0, \mathbf{\Sigma_x}), \tag{5.16}$$

where $\mathbf{\Sigma_x}$ is a self-adjoint HS operator satisfying

$$\langle \mathbf{\Sigma_x} f, f\rangle = \frac{\mathrm{Var}(\langle k(\mathbf{Y}, \cdot), f\rangle | \mathbf{X} = \mathbf{x})}{\mathrm{Var}(k(\mathbf{Y}, \cdot)|\mathbf{X} = \mathbf{x})} > 0 \tag{5.17}$$

for all $f \in \mathcal{H}$.

Proof. First, by the definition of σ_n, we have

$$\xi_n^0 := \sum_{i=1}^n \frac{s_n}{n\sigma_n} T_n(\mathbf{Z}_i) = \sum_{i=1}^n \frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}.$$

Define $\sigma_n^2(f) = \frac{s_n^2}{n}\mathrm{Var}(\langle T_n(\mathbf{Z}_1), f\rangle)$. Subsequently, we establish univariate convergence for all $f \in \mathcal{H}$:
Claim: For all $f \in \mathcal{H}$, we have $\sum_{i=1}^n \frac{s_n}{n\sigma_n(f)}\langle T_n(\mathbf{Z}_i), f\rangle \xrightarrow{D} N(0, \sigma(f)^2)$.
Proof:

Due to linearity, $\langle T_n(\mathbf{Z}_1), f\rangle$ is the first order approximation of a tree using the univariate response $f(\mathbf{Y}_i)$. Thus, it follows from Assumption **(F1)**–**(F5)** and **(D1)**–**(D7)** with the implications (5.8)–(5.10) and the arguments in the proof of Theorem 8 in Wager and Athey (2017) that

$$\sum_{i=1}^n \frac{s_n}{n\sigma_n(f)}\langle T_n(\mathbf{Z}_i), f\rangle \xrightarrow{D} N(0, 1). \tag{5.112}$$

From Theorem 5.3.4, we have

$$\frac{\sigma_n(f)}{\sigma_n} = \frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(T_n(\mathbf{Z}_1))} \to \sigma^2(f) > 0,$$

so that due to Slutsky's theorem,

$$\sum_{i=1}^n \frac{s_n}{n\sigma_n}\langle T_n(\mathbf{Z}_i), f\rangle = \frac{\sigma_n(f)}{\sigma_n}\sum_{i=1}^n \frac{s_n}{n\sigma_n(f)}\langle T_n(\mathbf{Z}_i), f\rangle \to N(0, \sigma^2(f)) \tag{5.113}$$

with $\sigma^2(f) > 0$. □

Now, we proof uniform tightness:
Claim: $(\xi_n^0)_{n\in\mathbb{N}}$ is uniformly tight.
Proof:

Because $\mathcal{H}$ is separable due to our assumptions on the kernel, there exists

a complete orthogonal basis $(e_j)_{j\in\mathbb{N}}$ of $\mathcal{H}$; see for instance Hsing and Eubank (2015). Let P_k be the projection operator onto the linear span of the first k elements of $(e_j)_{j\in\mathbb{N}}$, $S_k = \text{span}(e_1, \ldots, e_k)$. Because S_k is closed and linear, P_k is well defined. Moreover, for all $f \in \mathcal{H}$, we have $\langle f - P_k(f), P_k(f)\rangle = 0$. Furthermore, it can be shown that $P_k(f) = \sum_{j=1}^{k}\langle f, e_j\rangle e_j$.

We now verify condition (c) of Chen and White (1998, Lemma 3.2), which is a sufficient condition for tightness:

Claim: $\limsup_n \mathbb{E}[\|\xi_n^0 - P_k(\xi_n^0)\|_{\mathcal{H}}^2] \to 0$, as $k \to \infty$.

Proof: For any n, k, we have

$$\mathbb{E}[\|\xi_n^0 - P_k(\xi_n^0)\|_{\mathcal{H}}^2] = \mathbb{E}[\|\xi_n^0\|_{\mathcal{H}}^2] + \mathbb{E}[\|P_k(\xi_n^0)\|_{\mathcal{H}}^2] - 2\mathbb{E}[\langle \xi_n^0, P_k(\xi_n^0)\rangle_{\mathcal{H}}].$$

Furthermore, for all n, we have

$$\begin{aligned}\mathbb{E}[\|\xi_n^0\|_{\mathcal{H}}^2] &= \text{Var}(\xi_n^0) = \text{Var}\left(\sum_{i=1}^{n}\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\text{Var}(T_n(\mathbf{Z}_1))}}\right)\\ &= \frac{n}{n\text{Var}(T_n(\mathbf{Z}_1))}\text{Var}(T_n(\mathbf{Z}_1)) = 1.\end{aligned}$$

Because $P_k(\xi_n^0)$ is an orthogonal projection, we have

$$\mathbb{E}[\langle \xi_n^0, P_k(\xi_n^0)\rangle_{\mathcal{H}}] = \mathbb{E}[\|P_k(\xi_n^0)\|_{\mathcal{H}}^2].$$

Thus,

$$\mathbb{E}[\|\xi_n^0 - P_k(\xi_n^0)\|_{\mathcal{H}}^2] = 1 - \mathbb{E}[\|P_k(\xi_n^0)\|_{\mathcal{H}}^2].$$

Now for any k, we have

$$\begin{aligned}\mathbb{E}[\|P_k(\xi_n^0)\|_{\mathcal{H}}^2] &= \sum_{j=1}^{k}\mathbb{E}[\langle \xi_n^0, e_j\rangle_{\mathcal{H}}^2]\\ &= \frac{1}{\text{Var}(T_n(\mathbf{Z}_1))}\sum_{j=1}^{k}\mathbb{E}[\langle T_n(\mathbf{Z}_1), e_j\rangle_{\mathcal{H}}^2]\\ &= \sum_{j=1}^{k}\frac{\text{Var}(\langle T_n(\mathbf{Z}_1), e_j\rangle_{\mathcal{H}})}{\text{Var}(T_n(\mathbf{Z}_1))}\\ &\to \sum_{j=1}^{k}\frac{\text{Var}(\langle k(\mathbf{Y},\cdot), e_j\rangle|\mathbf{X}=\mathbf{x})}{\text{Var}(k(\mathbf{Y},\cdot)|\mathbf{X}=\mathbf{x})},\end{aligned}$$

as $n \to \infty$ due to (5.15) and the fact that the sum over k is finite. Additionally,

due to Hsing and Eubank (2015, Chapter 7), we have

$$\mathrm{Var}(k(\mathbf{Y},\cdot)|\mathbf{X}=\mathbf{x}) = \sum_{j=1}^{\infty} \mathrm{Var}(\langle k(\mathbf{Y},\cdot), e_j\rangle|\mathbf{X}=\mathbf{x}).$$

This means that

$$\begin{aligned}\limsup_n \mathbb{E}[\|\xi_n^0 - P_k(\xi_n^0)\|_{\mathcal{H}}^2] &= 1 - \liminf_n \mathbb{E}[\|P_k(\xi_n^0)\|_{\mathcal{H}}^2] \\ &= 1 - \sum_{j=1}^{k} \frac{\mathrm{Var}(\langle k(\mathbf{Y},\cdot), e_j\rangle|\mathbf{X}=\mathbf{x})}{\mathrm{Var}(k(\mathbf{Y},\cdot)|\mathbf{X}=\mathbf{x})} \\ &\to 0\end{aligned}$$

as $k \to \infty$. □

Consequently, $(\xi_n^0)_{n\in\mathbb{N}}$ is uniformly tight.

□

Univariate convergence together with tightness imply $\xi_n^0 \xrightarrow{D} N(0, \mathbf{\Sigma}_{\mathbf{x}})$; see for example Chen and White (1998, Lemma 3.1/3.2) or Hsing and Eubank (2015, Chapter 7). Since by Theorem 5.3.2 we have

$$\frac{1}{\sigma_n}(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})) = \xi_n^0 + o_p(1),$$

the result follows. □

Before being able to prove Theorem 5.3.6, we need a few preliminary results: Let in the following $\mathcal{H}^*$ be the dual space of $\mathcal{H}$, that is,

$$\mathcal{H}^* = \{F\colon \mathcal{H} \to \mathbb{R} \text{ linear, bounded, and continuous}\}.$$

Moreover, let

$$\mathcal{F} = \{F \in \mathcal{H}^*, \|F\|_{\mathcal{H}^*} \leq 1\}, \tag{5.114}$$

where $\|\cdot\|_{\mathcal{H}^*}$ is the operator norm on $\mathcal{H}^*$. Additionally, let $\ell^\infty(\mathcal{F})$ be the space of all bounded real-valued functions $\mathcal{F} \to \mathbb{R}$.

Due to the Riesz representation theorem, for each $F \in \mathcal{H}^*$ there exists exactly one $f_F \in \mathcal{H}$ such that $F(h) = \langle f_F, h\rangle$ for all $h \in \mathcal{H}$. Let us define the map $D\colon \mathcal{H} \to \ell^\infty(\mathcal{F})$ by

$$D(f)(F) = F(f) \text{ for } F \in \mathcal{F}. \tag{5.115}$$

Following the notation of empirical process theory, for $F \in \mathcal{F}$, we let

$$\mathbb{P}_{k,\mathbf{x}}F = D(\mu(\mathbf{x}))(F) = F(\mu(\mathbf{x})) = \mathbb{E}[F(k(\mathbf{Y}, \cdot)) \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[f_F(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}].$$

Thus, $\mathbb{P}_{k,\mathbf{x}}$ is the process associated with $k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}$ on $\mathcal{H}$. Similarly, let us for $F \in \mathcal{F}$ denote by $\hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}}$ the function defined by

$$(\hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}})F = \langle \hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}), f_F \rangle.$$

Moreover, define the Gaussian process $G_{\mathbb{P}_{k,\mathbf{x}}}$ on $\ell^\infty(\mathcal{F})$ by

$$G_{\mathbb{P}_{k,\mathbf{x}}}(F) = D(\xi)(F) = F(\xi) = \langle \xi, f_F \rangle_{\mathcal{H}},$$

where $\xi \sim N(0, \mathbf{\Sigma}_\mathbf{x})$ on $\mathcal{H}$, with $\mathbf{\Sigma}_\mathbf{x}$ as in Theorem 5.3.5.

González-Rodríguez and Colubi (2017) show that D is linear and continuous and that it has a continuous inverse. With this, it follows that:

Corollary 5.A.12. *For all n, $\frac{1}{\sigma_n}(\hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}}) \in \ell^\infty(\mathcal{F})$ and*

$$\frac{1}{\sigma_n}(\hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}}) \xrightarrow{D} G_{\mathbb{P}_{k,\mathbf{x}}}$$

in $\ell^\infty(\mathcal{F})$.

Proof. González-Rodríguez and Colubi (2017) show that D in (5.115) is a continuous bounded linear operator satisfying

$$\frac{1}{\sigma_n} D(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})) = D\left(\frac{1}{\sigma_n}(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))\right) \xrightarrow{D} D(\xi) = G_{\mathbb{P}_{k,\mathbf{x}}}$$

due to the continuous mapping theorem. Additionally, by the Riesz representation theorem,

$$\begin{aligned} D(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))(F) &= F(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})) = \langle f_F, \hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}) \rangle \\ &= (\hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}})F \end{aligned}$$

for all $F \in \mathcal{F}$, so that $D(\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x})) = \hat{\mathbb{P}}_{k,\mathbf{x}} - \mathbb{P}_{k,\mathbf{x}}$. □

This result enables us to use empirical process techniques, as we will do in the proof of Theorem 5.3.6. To prove Theorem 5.3.6, we start with the following important Lemma, in analogy to Kosorok (2003, Lemma 3):

Lemma 5.A.13. *Let Y_{ni}, $i = 1, \ldots, m_n, n \geq 1$ be a triangular array of mean zero independent (within rows) random variables. Let W_i, $i =$*

$1, \ldots, n$ *be i.i.d random variables, independent of* $(Y_{ni})_{n,i}$*, and with* $\mathbb{E}[W_i] = 0$ *and* $\mathrm{Var}(W_i) = 1$ *for all* i*. Additionally, assume that we have*

$$\sum_{i=1}^{m_n} \mathrm{Var}(Y_{ni}^2) \to \sigma_0 > 0 \tag{5.116}$$

and

$$\sum_{i=1}^{m_n} Y_{ni}^2 \xrightarrow{p} \sigma_0 > 0 \tag{5.117}$$

and moreover, for some $\delta > 0$*,*

$$\lim_{n\to\infty} \sum_{i=1}^{m_n} \frac{\mathbb{E}[|Y_{ni}|^{2+\delta}]}{\left(\sum_{j=1}^{m_n} \mathrm{Var}(Y_{nj})\right)^{1+\delta/2}} = 0. \tag{5.118}$$

Then, for $\mathbb{Y}_n = \{Y_1, \ldots, Y_{m_n}\}$*,*

$$\mathrm{Var}\left(\sum_{i=1}^{m_n} W_i Y_{ni} \,\middle|\, \mathbb{Y}_n\right) = \mathrm{Var}\left(\sum_{i=1}^{m_n} W_i Y_{ni} \,\middle|\, \mathbb{Y}_n\right) \xrightarrow{p} \sigma_0 \tag{5.119}$$

and

$$\sum_{i=1}^{m_n} \frac{\mathbb{E}[|W_i Y_{ni}|^{2+\delta} \mid \mathbb{Y}_n]}{\left(\sum_{j=1}^{m_n} \mathrm{Var}(W_j Y_{nj} \mid \mathbb{Y}_n)\right)^{1+\delta/2}} \xrightarrow{p} 0 \tag{5.120}$$

as $n \to \infty$*.*

Proof. First, by (5.117),

$$\mathrm{Var}\left(\sum_{i=1}^{m_n} W_i Y_{ni} \,\middle|\, \mathbb{Y}_n\right) = \sum_{i=1}^{m_n} \mathrm{Var}\left(W_i \mid \mathbb{Y}_n\right) Y_{ni}^2 = \sum_{i=1}^{m_n} Y_{ni}^2 \xrightarrow{p} \sigma_0,$$

which establishes (5.119). Similarly, we have

$$\sum_{i=1}^{m_n} \frac{\mathbb{E}[|W_i Y_{ni}|^{2+\delta} \mid \mathbb{Y}_n]}{\left(\sum_{j=1}^{m_n} \mathrm{Var}(W_j Y_{nj} \mid \mathbb{Y}_n)\right)^{1+\delta/2}} = \mathbb{E}[|W_1|^{2+\delta} \mid \mathbb{Y}_n] \sum_{i=1}^{m_n} \frac{|Y_{ni}|^{2+\delta}}{\left(\sum_{j=1}^{m_n} Y_{nj}^2\right)^{1+\delta/2}}$$

and

$$\sum_{i=1}^{m_n} \frac{|Y_{ni}|^{2+\delta}}{\left(\sum_{j=1}^{m_n} Y_{nj}^2\right)^{1+\delta/2}} = \sum_{i=1}^{m_n} \frac{|Y_{ni}|^{2+\delta}}{\left(\sum_{j=1}^{m_n} \mathrm{Var}(Y_{nj})\right)^{1+\delta/2}} \left(\frac{\sum_{j=1}^{m_n} \mathrm{Var}(Y_{nj})}{\sum_{j=1}^{m_n} Y_{nj}^2}\right)^{1+\delta/2}.$$

By Assumption (5.116) and (5.117), we have

$$\left(\frac{\sum_{i=1}^{m_n} \operatorname{Var}(Y_{ni})}{\sum_{j=1}^{m_n} Y_{ni}^2}\right)^{1+\delta/2} \xrightarrow{p} 1,$$

and, due to Markov's inequality and (5.118),

$$\mathbb{P}\left(\sum_{i=1}^{m_n} \frac{|Y_{ni}|^{2+\delta}}{\left(\sum_{j=1}^{m_n} \operatorname{Var}(Y_{nj})\right)^{1+\delta/2}} > \varepsilon\right) \leq \frac{1}{\varepsilon} \sum_{i=1}^{m_n} \frac{\mathbb{E}[|Y_{ni}|^{2+\delta}]}{\left(\sum_{j=1}^{m_n} \operatorname{Var}(Y_{nj})\right)^{1+\delta/2}} \to 0,$$

so that

$$\sum_{i=1}^{m_n} \frac{|Y_{ni}|^{2+\delta}}{\left(\sum_{j=1}^{m_n} \operatorname{Var}(Y_{nj})\right)^{1+\delta/2}} \cdot \frac{\left(\sum_{j=1}^{m_n} \operatorname{Var}(Y_{nj})\right)^{1+\delta/2}}{\left(\sum_{j=1}^{m_n} Y_{nj}^2\right)^{1+\delta/2}} = o_p(1),$$

which establishes the result. □

Theorem 5.3.6. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then,* (5.19) *holds.*

Proof. For this proof, we recall the definition of ξ_n in (5.7). For each subsample of size s_n of the data, we have a tree. For a given $\mathcal{S}$ we consider all such trees that are built using data points from $\mathcal{S}$. Thus, we consider the same "base" random forest built using all the data and select different trees depending on which subsample $\mathcal{S}$ we consider. Since s_n is of smaller order than n, $\mathbb{P}(|\mathcal{S}| \leq s_n) \to 0$, as $n \to \infty$. Thus, by the same arguments as in Athey et al. (2019, Theorem 5) combined with Theorem 5.3.2, we obtain

$$\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \mu(\mathbf{x}) = \frac{s_n}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} T_n(\mathbf{Z}_i) + o_p(\sigma_n) = \frac{s_n}{n} \sum_{i \in \mathcal{S}} \frac{n}{|\mathcal{S}|} T_n(\mathbf{Z}_i) + o_p(\sigma_n).$$

Due to $\sigma_n = \sqrt{s_n^2/n \cdot \operatorname{Var}(T_n(\mathbf{Z}_1))}$, we infer

$$\begin{aligned}\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \mu(\mathbf{x})\right) =& \frac{1}{\sqrt{n}} \sum_{i \in \mathcal{S}} \frac{n}{|\mathcal{S}|} \frac{T_n(\mathbf{Z}_i)}{\sqrt{\operatorname{Var}(T_n(\mathbf{Z}_1))}} + o_p(1) \\ =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{n}{|\mathcal{S}|} W_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{\operatorname{Var}(T_n(\mathbf{Z}_1))}} + o_p(1),\end{aligned}$$

with $(W_i)_{i=1}^n$ independent and $W_i \sim \text{Bernoulli}(1/2)$. Thus,

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right) = \frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \mu(\mathbf{x}) - (\hat{\mu}_n(\mathbf{x}) - \mu(\mathbf{x}))\right)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{n}{|\mathcal{S}|} W_i - 1 \right) \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} + o_p(1). \quad (5.121)$$

Recall our abbreviation

$$\xi_n^{\mathcal{S}} = \frac{1}{\sigma_n} \left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x}) \right)$$

from (5.18). Subsequently, we first prove the result for a simplified version of the sum in (5.121) consisting of independent summands. Let in the following

$$\tilde{W}_i = 2W_i - 1 \quad (5.122)$$

and

$$\xi_n^W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}. \quad (5.123)$$

Claim: It holds that

$$\xi_n^W \xrightarrow[W]{D} N(0, \mathbf{\Sigma_x}). \quad (5.124)$$

Proof:

The proof combines arguments from Kosorok (2003) with arguments made above and the equivalence of $\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})$ and a certain empirical process as in González-Rodríguez and Colubi (2017). Note that, since the W_i are i.i.d, $W_i \sim$ Bernoulli(1/2), $\mathbb{E}[\tilde{W}_i] = 0$ and $\mathrm{Var}(\tilde{W}_i) = 1$. First, we prove unconditional convergence:

Claim: It holds that

$$\xi_n^W \xrightarrow{D} N(0, \mathbf{\Sigma_x}). \quad (5.125)$$

Proof:

We start by verifying uniform tightness of the sequence $(\xi_n^W)_n$:

Claim: $\limsup_n \mathbb{E}[\|\xi_n^W - P_k(\xi_n^W)\|_{\mathcal{H}}^2] \to 0$ as $k \to \infty$.

Proof: For all n, we have

$$\begin{aligned}
&\mathbb{E}[\|\xi_n^W\|_{\mathcal{H}}^2] \\
=&\mathrm{Var}\left(\sum_{i=1}^{n} \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right) \\
=&\mathbb{E}\left[\mathrm{Var}\left(\sum_{i=1}^{n} \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}} \,\middle|\, \mathcal{Z}_n \right) \right] + \mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{n} \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}} \,\middle|\, \mathcal{Z}_n \right] \right).
\end{aligned}$$

For the first term in the above decomposition, we have

$$\mathbb{E}\left[\mathrm{Var}\left(\sum_{i=1}^{n}\tilde{W}_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}\;\middle|\;\mathcal{Z}_n\right)\right]=\mathbb{E}\left[\left\|\frac{T_n(\mathbf{Z}_i)}{\mathrm{Var}(T_n(\mathbf{Z}_1))}\right\|^2\right]=1.$$

And for the second term, we have

$$\begin{aligned}\mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{n}\tilde{W}_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}\;\middle|\;\mathcal{Z}_n\right]\right)&=\mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{n}\tilde{W}_i\;\middle|\;\mathcal{Z}_n\right]\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right)\\&=0.\end{aligned}$$

Thus, $\mathbb{E}[\|\xi_n^W\|_{\mathcal{H}}^2]=1=\mathrm{Var}(\xi_n)$. Similarly,

$$\begin{aligned}\mathbb{E}[\|P_k(\xi_n^W)\|_{\mathcal{H}}^2]&=\sum_{j=1}^{k}\mathbb{E}[\langle\xi_n^W,e_j\rangle_{\mathcal{H}}^2]\\&=\sum_{j=1}^{k}\frac{\mathrm{Var}(\tilde{W}_i\langle T_n(\mathbf{Z}_1),e_j\rangle_{\mathcal{H}})}{\mathrm{Var}(T_n(\mathbf{Z}_1))}\\&=\sum_{j=1}^{k}\frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1),e_j\rangle_{\mathcal{H}})}{\mathrm{Var}(T_n(\mathbf{Z}_1))}\end{aligned}$$

due to the same variance arguments, so that $\mathbb{E}[\|P_k(\xi_n^W)\|_{\mathcal{H}}^2]=\mathbb{E}[\|P_k(\xi_n)\|_{\mathcal{H}}^2]$ Thus, the claim follows by exactly the same argument as in the proof of Theorem 5.3.5. □

We now verify marginal convergence:

Claim: For all $f\in\mathcal{H}$, we have $\langle\xi_n^W,f\rangle\xrightarrow{D}N(0,\sigma^2(f))$, where $\sigma(f)>0$ is defined in Theorem 5.3.4.

Proof: We prove convergence using the Lyapunov central limit theorem similarly to Wager and Athey (2018, Theorem 8). First, with the arguments in the proof of Theorem 8 in Wager and Athey (2017), it can be shown that, under Assumption **(F1)**–**(F5)** and **(D1)**–**(D7)** with the implications (5.8)–(5.10), that we have

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{\mathbb{E}[|\langle T_n(\mathbf{Z}_i),f\rangle|^{2+\delta}]}{(n\mathrm{Var}(\langle T_n(\mathbf{Z}_1),f\rangle))^{1+\delta/2}}=0.$$

By Theorem 5.3.4, consequently also

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{\mathbb{E}[|\langle T_n(\mathbf{Z}_i),f\rangle|^{2+\delta}]}{(n\mathrm{Var}(T_n(\mathbf{Z}_1)))^{1+\delta/2}}=0,$$

that is, the Lyapunov condition holds for $\langle\xi_n,f\rangle$. As $\mathrm{Var}(\tilde{W}_1T_n(\mathbf{Z}_1))=1$ and

$\mathbb{E}[|\tilde{W}_i|^{2+\delta}] = \mathbb{E}[|\tilde{W}_1|^{2+\delta}] \leq 1$, we have

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{\mathbb{E}\left[\left|\tilde{W}_i \langle T_n(\mathbf{Z}_i), f\rangle\right|^{2+\delta}\right]}{\left(n\mathrm{Var}(\tilde{W}_i T_n(\mathbf{Z}_1))\right)^{1+\delta/2}} \leq \lim_{n\to\infty} \sum_{i=1}^{n} \frac{\mathbb{E}[|\langle T_n(\mathbf{Z}_i), f\rangle|^{2+\delta}]}{\left(n\mathrm{Var}(T_n(\mathbf{Z}_1))\right)^{1+\delta/2}} = 0,$$

so that the Lyapunov condition holds for $\langle \xi_n^W, f\rangle$. Finally, by the same arguments,

$$\mathrm{Var}(\langle \xi_n^W, f\rangle) = \frac{\mathrm{Var}(\langle T_n(\mathbf{Z}_1), f\rangle)}{\mathrm{Var}(T_n(\mathbf{Z}_1))} \to \sigma(f),$$

which shows the claim. □

Uniform tightness and convergence of univariate marginals together imply (5.125). □

Let us consider again the function D defined in (5.115) and the set $\mathcal{F} = \{F \in \mathcal{H}^* \colon \|F\|_{\mathcal{H}^*} \leq 1\}$ defined in (5.114). As mentioned above, $D \colon \mathcal{H} \to \ell^\infty(\mathcal{F})$ is continuous with a continuous inverse, and we consider the non-i.i.d empirical process

$$D(\xi_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D\left(\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right)$$

and similarly the multiplier process

$$D(\xi_n^W) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{W}_i D\left(\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right).$$

Using continuity of D, we showed $D(\xi_n^W) \xrightarrow{D} D(\xi)$, which in turn is a tight Gaussian element in $\ell^\infty(\mathcal{F})$; see González-Rodríguez and Colubi (2017).

Having shown unconditional convergence, we show conditional convergence of finite-dimensional marginals of $D(\xi_n^W)$:
Claim: For all $K \in \mathbb{N}$ and $(f_1, \ldots, f_K) \in \mathcal{F}^K$,

$$\left(D(\xi_n^W)(f_1), \ldots, D(\xi_n^W)(f_K)\right) \xrightarrow[W]{D} \left(D(\xi)(f_1), \ldots, D(\xi)(f_K)\right). \tag{5.126}$$

Proof: By the Cramer-Wold device, it suffices to show

$$\left(D(\xi_n^W)(f_1), \ldots, D(\xi_n^W)(f_K)\right) \cdot \mathbf{w} \xrightarrow[W]{D} \left(D(\xi)(f_1), \ldots, D(\xi)(f_K)\right) \cdot \mathbf{w}, \tag{5.127}$$

for any $\mathbf{w} \in \mathbb{R}^K$. This in turn is implied if for all $F \colon \ell^\infty(\mathcal{F}) \to \mathbb{R}$ linear and

continuous, it holds that

$$F(D(\xi_n^W)) \xrightarrow[W]{D} F(D(\xi)) \tag{5.128}$$

because $F_K \colon \ell^\infty(\mathcal{F}) \to \mathbb{R}$, $F_K(D(\xi)) = (D(\xi)(f_1), \ldots, D(\xi)(f_K)) \cdot \mathbf{w}$ is linear and continuous. Consider a linear and continuous function $F \colon \mathcal{H} \to \mathbb{R}$. Because $F \circ D \colon \mathcal{H} \to \mathbb{R}$ is linear and continuous from $\mathcal{H}$ to $\mathbb{R}$, by the Riesz representation theorem, we have

$$\begin{aligned} F(D(\xi_n^W)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{W}_i F \circ D \left(\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{W}_i \frac{\langle T_n(\mathbf{Z}_i), f_F \rangle}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}, \end{aligned}$$

for a unique $f_F \in \mathcal{H}$. Combining the arguments to prove the Lyapunov conditions in Wager and Athey (2018, Theorem 8) with Theorem 5.3.4, we see that conditions (5.116) and (5.118) of Lemma 5.A.13 hold for $Y_{ni} = \frac{\langle T_n(\mathbf{Z}_i), f_F \rangle}{\sqrt{n \mathrm{Var}(T_n(\mathbf{Z}_1))}}$. Similarly, Wager and Athey (2018, Lemma 12) implies that (5.117) holds as well for Y_{ni}. Since $(\tilde{W}_i)_i$ is i.i.d. with expectation 0 and variance 1, it follows from Lemma 5.A.13 that the Lyapunov condition for $\tilde{W}_i Y_{ni}$ holds in probability, that is, (5.119) and (5.120) hold. Thus, we can find for any subsequence a further subsequence indexed by say l such that Lyapunov condition for $\sum_i \tilde{W} Y_{li}$ given $\mathcal{Z}_l$ hold almost surely. Arguing pointwise for fixed $\mathcal{Z}_l$ implies

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left| \mathbb{E}[h(F(D(\xi_l^{\mathcal{S}}))) \mid \mathcal{Z}_l] - \mathbb{E}[h(F(D(\xi)))] \right| \to 0 \text{ a.s.;}$$

see Kosorok (2003). Using an argument by contradiction as in Ćevid et al. (2022, Lemma 14), this in turn means

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left| \mathbb{E}[h(F(D(\xi_l^{\mathcal{S}}))) \mid \mathcal{Z}_l] - \mathbb{E}[h(F(D(\xi)))] \right| \xrightarrow{p} 0,$$

proving the claim. □

Combining unconditional convergence given in (5.125) and conditional finite-dimensional convergence given in (5.126) with the arguments in Kosorok (2003, Theorem 2) then gives

$$D(\xi_n^W) \xrightarrow[W]{D} D(\xi). \tag{5.129}$$

Finally, due to continuity of the inverse of D, this implies (5.124).

□

Having shown (5.124), it holds that

Claim:

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{n}{|\mathcal{S}|}W_i-1\right)\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(2W_i-1)\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\\&=\left(\frac{n}{|\mathcal{S}|}-2\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\\&\xrightarrow{p}0.\end{aligned}\tag{5.130}$$

Proof:

Indeed, $\left(\frac{n}{|\mathcal{S}|}-2\right)=o_p(1)$, and due to

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=1}^{n}W_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right\|>\varepsilon\right)&\leq\frac{1}{\varepsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n}W_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{n\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right)\\&=\frac{1}{\varepsilon^2}\frac{\mathrm{Var}(W_1T_n(\mathbf{Z}_1))}{\mathrm{Var}(T_n(\mathbf{Z}_1))}\end{aligned}$$

with

$$\frac{\mathrm{Var}(W_1T_n(\mathbf{Z}_1))}{\mathrm{Var}(T_n(\mathbf{Z}_1))}=\frac{1/4\mathrm{Var}(T_n(\mathbf{Z}_1))+1/4\mathrm{Var}(T_n(\mathbf{Z}_1))}{\mathrm{Var}(T_n(\mathbf{Z}_1))}=\frac{1}{2}<\infty,$$

we have

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i\frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right\|=\mathcal{O}_p(1),$$

establishing (5.130).

□

Thus, we have (5.124), that is, $\xi_n^W\xrightarrow[W]{D}N(0,\mathbf{\Sigma_x})$. Moreover, we have

$$\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})-\hat{\mu}_n(\mathbf{x})\right)=\xi_n^W+o_p(1),$$

by combining (5.121) with (5.130). Let us denote as in the main text $\xi_n^{\mathcal{S}}=\frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})-\hat{\mu}_n(\mathbf{x})\right)$, and let $\mathcal{D}_n$ be the difference

$$\mathcal{D}_n=\xi_n^{\mathcal{S}}-\xi_n^W,$$

so that $\|\mathcal{D}_n\|_{\mathcal{H}}=o_p(1)$. With this, we can finally show that (5.20) holds, that

is,

$$\sup_{h\in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}\left[h\left(\xi_n^{\mathcal{S}}\right) \mid \mathcal{Z}_n\right] - \mathbb{E}[h(\xi)]\right| \xrightarrow{p} 0.$$

Indeed, we have

$$\begin{aligned}
&\sup_{h\in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}\left[h\left(\xi_n^{\mathcal{S}}\right) \mid \mathcal{Z}_n\right] - \mathbb{E}[h(\xi)]\right| \\
&\leq \sup_{h\in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}\left[h\left(\xi_n^{\mathcal{S}}\right) \mid \mathcal{Z}_n\right] - \mathbb{E}[h(\xi_n^W) \mid \mathcal{Z}_n]\right| + \sup_{h\in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}[h(\xi_n^W) \mid \mathcal{Z}_n] - \mathbb{E}[h(\xi)]\right|.
\end{aligned} \tag{5.131}$$

The second term goes to zero in probability by (5.124), and the first term satisfies

$$\begin{aligned}
\sup_{h\in \mathrm{BL}_1(\mathcal{H})} \left|\mathbb{E}\left[h\left(\xi_n^{\mathcal{S}}\right) \mid \mathcal{Z}_n\right] - \mathbb{E}[h(\xi_n^W) \mid \mathcal{Z}_n]\right| &\leq \sup_{h\in \mathrm{BL}_1(\mathcal{H})} \mathbb{E}[|h(\xi_n^{\mathcal{S}}) - h(\xi_n^W)| \mid \mathcal{Z}_n] \\
&\leq \mathbb{E}[\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2) \mid \mathcal{Z}_n]
\end{aligned}$$

because for all $h \in \mathrm{BL}_1(\mathcal{H})$, h is Lipschitz with constant bounded by 1, and $|h(f_1) - h(f_2)| \leq 2\sup_{f\in\mathcal{H}} |h(f)| \leq 2$. Moreover, since $(\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2))_n$ is a bounded sequence, it is uniformly integrable; see Dudley (2002, Chapter 10.3). It follows by an extension of the Dominated Convergence Theorem for convergence in probability (Dudley, 2002, Theorem 10.3.6) that $\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2) = o_p(1)$, which implies $\mathbb{E}[\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2)] \to 0$. Since $\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2)$ is also nonnegative and

$$o(1) = \mathbb{E}[\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2)] = \mathbb{E}[\mathbb{E}[\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2) \mid \mathcal{Z}_n]],$$

this implies that $\mathbb{E}[\min(\|\mathcal{D}_n\|_{\mathcal{H}}, 2) \mid \mathcal{Z}_n] \xrightarrow{p} 0$. This convergence, together with the above bound, shows that the first part of (5.131) also goes to zero in probability.

□

Finally, we show that the variance of finite dimensional marginals can be estimated consistently:

Corollary 5.3.7. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then, for any $F\colon \mathcal{H} \to \mathbb{R}^q$ linear and continuous,*

$$\mathbb{E}\left[\frac{1}{\sigma_n^2}\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)^\top \,\middle|\, \mathcal{Z}_n\right] \xrightarrow{p} F \circ \mathbf{\Sigma}_{\mathbf{x}}. \tag{5.21}$$

Proof. Define

$$F \circ \hat{\mathbf{\Sigma}}_n = \mathbb{E}\left[\frac{1}{\sigma_n^2}\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)\left(F(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x})) - F(\hat{\mu}_n(\mathbf{x}))\right)^\top \,\middle|\, \mathcal{Z}_n\right], \tag{5.132}$$

and note that $F \circ \hat{\mathbf{\Sigma}}_n = \mathbb{E}\left[F(\xi_n^{\mathcal{S}})F(\xi_n^{\mathcal{S}})^\top \mid \mathcal{Z}_n\right]$. Similarly, we define

$$F \circ \hat{\mathbf{\Sigma}}_n^o = \mathbb{E}\left[F(\xi_n^W)F(\xi_n^W)^\top \,\middle|\, \mathcal{Z}_n\right], \tag{5.133}$$

with ξ_n^W defined as in (5.123). We will first show in several steps that:
Claim: For all $\mathbf{w} \in \mathbb{R}^q$, we have

$$\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n)\mathbf{w} \xrightarrow{p} \mathbf{w}^\top(F \circ \mathbf{\Sigma}_{\mathbf{x}})\mathbf{w}. \tag{5.134}$$

Proof:

To proof the claim, we first show:
Claim: For all $\mathbf{w} \in \mathbb{R}^q$, we have

$$\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n^o)\mathbf{w} \xrightarrow{p} \mathbf{w}^\top(F \circ \mathbf{\Sigma}_{\mathbf{x}})\mathbf{w}. \tag{5.135}$$

Proof: First, note that we may define $F_{\mathbf{w}} \in \mathcal{H}^*$ by $F_{\mathbf{w}}(f) = \mathbf{w}^\top F(f)$. Particularly, it is linear, and $\|F_{\mathbf{w}}(f_1) - F_{\mathbf{w}}(f_2)\| \leq \|\mathbf{w}\|_{\mathbb{R}^q}\|F(f_1) - F(f_2)\|_{\mathbb{R}^q}$, so that it is also continuous. Then, we have

$$\begin{aligned}\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n^o)\mathbf{w} &= \mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{W}_i \frac{F_{\mathbf{w}} \circ T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}\right\|^2 \,\middle|\, \mathcal{Z}_n\right] \\ &= \frac{1}{n}\sum_{i=1}^n \frac{(F_{\mathbf{w}} \circ T_n(\mathbf{Z}_i))^2}{\mathrm{Var}(T_n(\mathbf{Z}_1))},\end{aligned}$$

because $\mathbb{E}[\tilde{W}_i^2] = 1$ and because the cross-terms are of the form

$$\mathbb{E}[\tilde{W}_i\tilde{W}_j]\frac{F_{\mathbf{w}} \circ T_n^2(\mathbf{Z}_i) \cdot F_{\mathbf{w}} \circ T_n^2(\mathbf{Z}_j)}{\mathrm{Var}(T_n(\mathbf{Z}_1))} = 0.$$

As already argued in the proof of Theorem 5.3.6, under Assumption **(F1)**–**(F5)** and **(D1)**–**(D7)**, the arguments in the proof of Lemma 12 in Wager and Athey (2017) imply that

$$\frac{1}{n}\sum_{i=1}^n \frac{(F_{\mathbf{w}} \circ T_n(\mathbf{Z}_i))^2}{\mathrm{Var}(T_n(\mathbf{Z}_1))} = \frac{1}{n}\sum_{i=1}^n \frac{\langle f_{\mathbf{w}}, T_n(\mathbf{Z}_i)\rangle^2}{\mathrm{Var}(T_n(\mathbf{Z}_1))} \xrightarrow{p} \sigma^2(f_{\mathbf{w}})$$

for the unique $f_{\mathbf{w}} \in \mathcal{H}$ given by the Riesz representation theorem. Moreover, by consistency arguments, we have $\sigma^2(f_{\mathbf{w}}) = \mathbf{w}^\top (F \circ \mathbf{\Sigma}_{\mathbf{x}})\mathbf{w}$, proving the claim. □

In the proof of Theorem 5.3.6, we showed $\xi_n^{\mathcal{S}} = \xi_n^W + o_p(1)$. To show that (5.134) follows from (5.135), we now strengthen this to:

Claim:

$$\mathbb{E}[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}^2] = o(1). \tag{5.136}$$

Proof: We recall the argument in the beginning of Theorem 5.3.6. By construction, we always consider the same forest and just use different trees or subsamples for each $\mathcal{S}$, namely such that the subset of size s_n is included in $\mathcal{S}$. Since s_n is of smaller order than n, $\mathbb{P}(|\mathcal{S}| \leq s_n) \to 0$, as $n \to \infty$. Thus, by the same arguments as in Athey et al. (2019, Theorem 5) combined with the claim (5.88), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sigma_n}\left((\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \mathbb{E}[\hat{\mu}_n(\mathbf{x})]) - \frac{s_n}{n}\sum_{i\in\mathcal{S}}\frac{n}{|\mathcal{S}|}T_n(\mathbf{Z}_i)\right)\right\|_{\mathcal{H}}^2 \mathbb{1}\{|\mathcal{S}| > s_n\}\right] \to 0.$$

Moreover, using that we have

$$\frac{\|\mathbb{E}[\hat{\mu}(\mathbf{x})] - \mu(\mathbf{x})\|_{\mathcal{H}}}{\sigma_n} = o(1),$$

as shown in the proof of Theorem 5.3.2, this convergence also holds with $\mathbb{E}[\hat{\mu}(\mathbf{x})]$ replaced by $\mu(\mathbf{x})$, or

$$\mathbb{E}\left[\left\|\xi_n^{\mathcal{S}} - \xi_n^W\right\|_{\mathcal{H}}^2 \mathbb{1}\{|\mathcal{S}| > s_n\}\right] \to 0.$$

In the case $|\mathcal{S}| \leq s_n$, we set

$$\sum_{i\in\mathcal{S}}\frac{n}{|\mathcal{S}|}T_n(\mathbf{Z}_i) = \hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) = 0 \in \mathcal{H}$$

to zero. Then, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\|\xi_n^{\mathcal{S}} - \xi_n^W\right\|_{\mathcal{H}}^2\right] - \mathbb{E}\left[\left\|\xi_n^{\mathcal{S}} - \xi_n^W\right\|_{\mathcal{H}}^2 \mathbb{1}\{|\mathcal{S}| > s_n\}\right]\right| \\
=&\mathbb{E}\left[\left\|\xi_n^{\mathcal{S}} - \xi_n^W\right\|_{\mathcal{H}}^2 \mathbb{1}\{|\mathcal{S}| \leq s_n\}\right] \\
=&\|\mu(\mathbf{x})\|_{\mathcal{H}}^2 \frac{\mathbb{P}(|\mathcal{S}| \leq s_n)}{\sigma_n^2}.
\end{aligned}$$

From the proof of Theorem 5.3.5 and the fact that $s_n = n^\beta$ with $\beta < 1$, it follows that

$$\sigma_n^2 = \Omega\left(\frac{s_n}{n\log(s_n)^p}\right) = \Omega(n^{\beta-(1+\varepsilon)})$$

for $\varepsilon > 0$ arbitrarily small. On the other hand, we can employ a Hoeffding bound on $\mathbb{P}(|\mathcal{S}| \leq s_n)$ to obtain

$$\mathbb{P}(|\mathcal{S}| \leq s_n) = \mathbb{P}(|\mathcal{S}| - n/2 \leq s_n - n/2) \leq \frac{\mathrm{Var}(|\mathcal{S}|)}{(s_n - n/2)^2} = \frac{n/4}{s_n^2 + n^2/4 - s_n n},$$

so that

$$\mathbb{P}(|\mathcal{S}| \leq s_n) \leq \frac{1}{4}\frac{1}{n/4 + n^{2\beta-1} - n^\beta} = \mathcal{O}\left(\frac{1}{n}\right).$$

This results in

$$\frac{\mathbb{P}(|\mathcal{S}| \leq s_n)}{\sigma_n^2} = \mathcal{O}\left(n^{1+\varepsilon-\beta-1}\right) = \mathcal{O}\left(n^{\varepsilon-\beta}\right).$$

Since ε can be chosen arbitrarily small, this converges to 0.

□

Having shown (5.136), we have that

$$\begin{aligned}
&\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n)\mathbf{w} \\
=&\mathbb{E}[\mathbf{w}^\top(F(\xi_n^W) + F(D_n))(F(\xi_n^W) + F(D_n))^\top\mathbf{w} \mid \mathcal{Z}_n] \\
=&\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n^o)\mathbf{w} + \mathbb{E}[\mathbf{w}^\top F(D_n)F(D_n)^\top\mathbf{w} \mid \mathcal{Z}_n] + 2\mathbb{E}[\mathbf{w}^\top F(\xi_n^W)F(D_n)^\top\mathbf{w} \mid \mathcal{Z}_n],
\end{aligned}$$

where $D_n = \xi_n^{\mathcal{S}} - \xi_n^W$. Note that $\mathbb{E}[\|D_n\|_{\mathcal{H}}^2] = \mathbb{E}[\mathbb{E}[\|D_n\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n]] = o(1)$ implies that $\mathbb{E}[\|D_n\|_{\mathcal{H}}^2 \mid \mathcal{Z}] = o_p(1)$; see Durrett (1996, Lemma 2.2.2). Moreover,

$$\begin{aligned}
\mathbb{E}[\mathbf{w}^\top F(D_n)F(D_n)^\top\mathbf{w} \mid \mathcal{Z}] &= \mathbb{E}[|\mathbf{w}^\top F(D_n)|^2 \mid \mathcal{Z}] \\
&\leq \|F_{\mathbf{w}}\|_{\mathcal{H}^*}^2 \mathbb{E}[\|D_n\|_{\mathcal{H}}^2 \mid \mathcal{Z}] \\
&= o_p(1)
\end{aligned}$$

by (5.136). Similarly, by Hölder's inequality,

$$\begin{aligned}
\mathbb{E}[\mathbf{w}^\top F(\xi_n^W)F(D_n)^\top\mathbf{w} \mid \mathcal{Z}_n] &\leq \mathbb{E}[|\mathbf{w}^\top F(\xi_n^W)|^2 \mid \mathcal{Z}_n]^{1/2} \cdot \mathbb{E}[|\mathbf{w}^\top F(D_n)|^2 \mid \mathcal{Z}_n]^{1/2} \\
&\xrightarrow{p} 0.
\end{aligned}$$

Thus, $\left|\mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n)\mathbf{w} - \mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n^o)\mathbf{w}\right| \xrightarrow{p} 0$, which shows (5.134).

□

Finally, (5.134) implies the result. Indeed, for a matrix $\mathbf{A} \in \mathbb{R}^{q\times q}$, define the operator $\mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{q^2}$ that concatenates the rows of $\mathbf{A}$ on top of each other. This operator is continuous and invertible with a continuous inverse. Moreover, for any $\mathbf{w}$, we can consider the element $\tilde{\mathbf{w}} = \mathbf{w} \otimes \mathbf{w}^\top \in \mathbb{R}^{q^2}$ satisfying $\mathbf{w}^\top \mathbf{A}\mathbf{w} = \tilde{\mathbf{w}}^\top \mathrm{vec}(\mathbf{A})$ such that we have

$$\tilde{\mathbf{w}}^\top \mathrm{vec}(F \circ \hat{\mathbf{\Sigma}}_n) = \mathbf{w}^\top(F \circ \hat{\mathbf{\Sigma}}_n)\mathbf{w} \xrightarrow{p} \mathbf{w}^\top(F \circ \mathbf{\Sigma}_{\mathbf{x}})\mathbf{w} = \tilde{\mathbf{w}}^\top \mathrm{vec}(F \circ \mathbf{\Sigma}_{\mathbf{x}}).$$

Utilizing the Cramer-Wold device and the fact that convergence in distribution to a constant is equivalent to convergence in probability, this implies that $\mathrm{vec}(F \circ \hat{\mathbf{\Sigma}}_n) \xrightarrow{p} \mathrm{vec}(F \circ \mathbf{\Sigma}_{\mathbf{x}})$. By continuity of the inverse of the vec operator, this implies the result.

□

Corollary 5.3.8. *Assume conditions **(F1)**–**(F5)**, **(D1)**–**(D7)**, **(K1)**, and **(K2)** hold. Then,*

$$\frac{\mathbb{E}[\|\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n]}{\sigma_n^2} \xrightarrow{p} 1. \tag{5.22}$$

Proof. First, we have

$$\begin{aligned}&\frac{\mathbb{E}[\|\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n]}{\sigma_n^2}\\ =&\mathbb{E}\left[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n\right] + \mathbb{E}\left[\|\xi_n^W\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n\right] + 2\mathbb{E}[\langle \xi_n^{\mathcal{S}} - \xi_n^W, \xi_n^W\rangle_{\mathcal{H}} \mid \mathcal{Z}_n],\end{aligned}$$

where we recall

$$\begin{aligned}\xi_n^{\mathcal{S}} &= \frac{1}{\sigma_n}\left(\hat{\mu}_n^{\mathcal{S}}(\mathbf{x}) - \hat{\mu}_n(\mathbf{x})\right),\\ \xi_n^W &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}.\end{aligned}$$

As we proved in Corollary 5.3.7, as a consequence of (5.136), we have

$$\mathbb{E}\left[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n\right] = o_p(1).$$

Moreover, using Cauchy–Schwarz inequality and Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}[|\langle \xi_n^{\mathcal{S}} - \xi_n^W, \xi_n^W\rangle_{\mathcal{H}}|] &\le \mathbb{E}[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}\|\xi_n^W\|_{\mathcal{H}}]\\ &\le \mathbb{E}[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}^2]^{1/2}\mathbb{E}[\|\xi_n^W\|_{\mathcal{H}}^2]^{1/2}.\end{aligned}$$

Recall that we argued $\mathbb{E}[\|\xi_n^W\|_{\mathcal{H}}^2] = 1$, and $\mathbb{E}[\|\xi_n^{\mathcal{S}} - \xi_n^W\|_{\mathcal{H}}^2] = o(1)$ above. This thus implies $\mathbb{E}[|\langle \xi_n^{\mathcal{S}} - \xi_n^W, \xi_n^W \rangle_{\mathcal{H}}|] = o(1)$, which in turn implies

$$|\mathbb{E}[\langle \xi_n^{\mathcal{S}} - \xi_n^W, \xi_n^W \rangle_{\mathcal{H}} \mid \mathcal{Z}_n]| \leq \mathbb{E}[|\langle \xi_n^{\mathcal{S}} - \xi_n^W, \xi_n^W \rangle_{\mathcal{H}}| \mid \mathcal{Z}_n] = o_p(1).$$

Thus, it remains to show:

Claim: $\mathbb{E}\left[\|\xi_n^W\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n\right] \xrightarrow{p} 1$.

Proof: First, note that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}, \tilde{W}_j \frac{T_n(\mathbf{Z}_j)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right\rangle_{\mathcal{H}} \middle| \mathcal{Z}_n\right] \\
=&\mathbb{E}[\tilde{W}_i \tilde{W}_j] \left\langle \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}}, \frac{T_n(\mathbf{Z}_j)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right\rangle_{\mathcal{H}} \\
=&0.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\|\xi_n^W\|_{\mathcal{H}}^2 \mid \mathcal{Z}_n\right] &= \mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{W}_i \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right\|_{\mathcal{H}}^2 \middle| \mathcal{Z}_n\right] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\tilde{W}_i^2 \mid \mathcal{Z}_n] \left\| \frac{T_n(\mathbf{Z}_i)}{\sqrt{\mathrm{Var}(T_n(\mathbf{Z}_1))}} \right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n\mathrm{Var}(T_n(\mathbf{Z}_1))} \sum_{i=1}^n \|T_n(\mathbf{Z}_i)\|_{\mathcal{H}}^2.
\end{aligned}$$

Thus, we need to show that

$$\frac{\frac{1}{n}\sum_{i=1}^n \|T_n(\mathbf{Z}_i)\|_{\mathcal{H}}^2}{\mathrm{Var}(T_n(\mathbf{Z}_1))} \xrightarrow{p} 1.$$

But due to assumption (**D3**), this can be shown using the same steps as at the end of the proof of Lemma 12 in Wager and Athey (2017), with $\|T_n(\mathbf{Z}_i)\|_{\mathcal{H}}^2$ in place of their $T_1^2(Z_i)$.

□

□

Corollary 5.4.1. *Assume conditions (**F1**)–(**F5**) and (**D1**)–(**D7**) for both groups, (**K1**), and (**K2**) hold, together with strong ignorability. Also*

assume that $n_0, n_1 \to \infty$ with $n_0/n_1 \to 1$. Then, for $\mathcal{S}_0$, $\mathcal{S}_1$ independent,

$$\left\| \frac{1}{\sigma_{n_1,1}}(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}}(\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow[W]{D} \|\xi_0 - \xi_1\|^2_{\mathcal{H}} \tag{5.26}$$

and

$$\left\| \frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow{D} \|\xi_0 - \xi_1\|^2_{\mathcal{H}}. \tag{5.27}$$

Moreover, if the ratio $\sigma_{n_0,0}/\sigma_{n_1,1}$ converges to some real number $c_2(\mathbf{x})$ that is bounded away from 0 and ∞ as the sample sizes n_0, n_1 tend to infinity, we obtain

$$\frac{1}{\sigma^2_{n_1,1}} \left\| (\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - (\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \right\|^2_{\mathcal{H}} \xrightarrow[W]{D} \|\xi_0 - c_2(\mathbf{x})\xi_1\|^2_{\mathcal{H}} \tag{5.28}$$

and

$$\frac{1}{\sigma^2_{n_1,1}} \|(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - (\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x}))\|^2_{\mathcal{H}} \xrightarrow{D} \|\xi_0 - c_2(\mathbf{x})\xi_1\|^2_{\mathcal{H}}. \tag{5.29}$$

Proof. Using independence of $\hat{\mu}_{n_1,1}(\mathbf{x})$ and $\hat{\mu}_{n_0,0}(\mathbf{x})$ for all n_0, n_1, together with Theorem 5.3.5, it follows that

$$\frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \xrightarrow{D} \xi_1 - \xi_0. \tag{5.137}$$

Similarly, due to independence of $\mathcal{S}_0$ and $\mathcal{S}_1$, the arguments in the proof of Theorem 5.3.6 can be repeated to obtain

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left| \mathbb{E}\left[h\left(\frac{1}{\sigma_{n_1,1}}(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}}(\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \right) \middle| \mathcal{Z}_{n_{01}} \right] - \mathbb{E}[h(\xi_1 - \xi_0)] \right| \xrightarrow{p} 0,$$

or in other words

$$\frac{1}{\sigma_{n_1,1}}(\hat{\mu}^{\mathcal{S}_1}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - \frac{1}{\sigma_{n_0,0}}(\hat{\mu}^{\mathcal{S}_0}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \xrightarrow[W]{D} \xi_1 - \xi_0. \tag{5.138}$$

Finally, if (5.30) holds, it follows from the arguments in the proof of Theorem 5.3.4 that, ignoring smaller order terms,

$$\begin{aligned}\frac{\sigma_{n_0,0}^2}{\sigma_{n_1,1}^2} &= \frac{\mathrm{Var}(\mathbb{E}[\frac{1}{N_{\mathbf{x}}^0}\mathbb{1}\{\mathbf{X}_2 \in \mathcal{L}^0(\mathbf{x})\} \mid \mathbf{X}_1])\,\mathrm{Var}(k(\mathbf{Y}^0,\cdot) \mid \mathbf{X}=\mathbf{x})}{\mathrm{Var}(\mathbb{E}[\frac{1}{N_{\mathbf{x}}^1}\mathbb{1}\{\mathbf{X}_2 \in \mathcal{L}^1(\mathbf{x})\} \mid \mathbf{X}_1])\,\mathrm{Var}(k(\mathbf{Y}^1,\cdot) \mid \mathbf{X}=\mathbf{x})} \\ &\to c(\mathbf{x})\frac{\mathrm{Var}(k(\mathbf{Y}^0,\cdot) \mid \mathbf{X}=\mathbf{x})}{\mathrm{Var}(k(\mathbf{Y}^1,\cdot) \mid \mathbf{X}=\mathbf{x})} = c_2(\mathbf{x})\end{aligned}$$

where $c(\mathbf{x})$ is as in (5.30). It thus follows from Slutsky's theorem that

$$\begin{aligned}&\frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \\ =&\frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{\sigma_{n_0,0}}{\sigma_{n_1,1}}\frac{1}{\sigma_{n_0,0}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})) \\ \xrightarrow{D}&\xi_1 - c_2(\mathbf{x})\xi_0. \end{aligned} \tag{5.139}$$

and similarly

$$\frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_1,1}^{\mathcal{S}_1}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})) - \frac{1}{\sigma_{n_1,1}}(\hat{\mu}_{n_0,0}^{\mathcal{S}_0}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \xrightarrow[W]{D} \xi_1 - c_2(\mathbf{x})\xi_0. \tag{5.140}$$

Since $f \mapsto \|f\|_{\mathcal{H}}^2$ is continuous, (5.26)–(5.29) follow from (5.137)–(5.140) combined with the continuous mapping theorem.

□

Theorem 5.4.2. *Assume conditions **(F1)**–**(F5)** and **(D1)**–**(D7)** for both groups, **(K1)**–**(K3)** hold, together with strong ignorability and* (5.30)*. Then, as $n_0, n_1 \to \infty$ such that $n_0/n_1 \to 1$,*

(i) φ has a valid type-I error. That is, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$,

$$\limsup_{n_0,n_1} \mathbb{P}\left(\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n_1,\alpha}\right) \le \alpha.$$

(ii) φ has power going to 1. That is, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \neq \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$,

$$\lim_{n_0,n_1} \mathbb{P}\left(\frac{1}{\sigma_{n_1,1}^2}\|\hat{\mu}_{n_0,0}(\mathbf{x}) - \hat{\mu}_{n_1,1}(\mathbf{x})\|_{\mathcal{H}}^2 > c_{n_1,\alpha}\right) = 1.$$

Proof. To simplify the proof, we assume $n_0 = n_1 = n$. Since we assume that $n_0/n_1 \to 1$, this will not impact our asymptotic results. Let in the following

for $j \in \{0,1\}$

$$\xi_{n,j}^{\mathcal{S}_j} = \frac{1}{\sigma_{n,1}} \left(\hat{\mu}_{n,j}^{\mathcal{S}_j}(\mathbf{x}) - \hat{\mu}_{n,j}(\mathbf{x}) \right),$$

where we emphasize the fixed 1 in $\sigma_{n,1}$. We first note that by (5.28), the sequence $\|\xi_{n,1}^{\mathcal{S}_1} - \xi_{n,0}^{\mathcal{S}_0}\|_{\mathcal{H}}^2$, $n \in \mathbb{N}$, is uniformly tight, which in turn implies that there exists a large enough number $M_\alpha < \infty$ such that we have

$$\sup_n \mathbb{P}\left(\|\xi_{n,1}^{\mathcal{S}_1} - \xi_{n,0}^{\mathcal{S}_0}\|_{\mathcal{H}}^2 > M_\alpha \right) \leq \alpha.$$

Since for each n, $c_{n,\alpha}$ is the smallest value such that (5.33) holds, we have $c_{n,\alpha} \leq M_\alpha < \infty$ for all n. In particular, $\sup_n c_{n,\alpha} \leq M_\alpha < \infty$, and $c_{n,\alpha}$ is a bounded sequence in $\mathbb{R}$. This allows us to find a convergent subsequence below. Second, we note that the distributions of $\|\xi_0\|_{\mathcal{H}}^2$ and $\|\xi_1\|_{\mathcal{H}}^2$ are dominated by the Lebesgue measure. Consequently,

$$z \mapsto \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2 > z)$$

is continuous. This in particular means that, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$,

$$\mathbb{P}\left(\frac{1}{\sigma_{n,1}^2} \|\hat{\mu}_{n,1}(\mathbf{x}) - \hat{\mu}_{n,0}(\mathbf{x})\|_{\mathcal{H}}^2 > z \right) \to \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2 > z)$$

for all $z \geq 0$ by (5.27). We consider an arbitrary subsequence of n, $n(\ell)$. By (5.26), a further subsequence $m = n(\ell(m))$ can be chosen such that

$$\sup_{h \in \mathrm{BL}_1(\mathcal{H})} \left| \mathbb{E}[h(\|\xi_{m,1}^{\mathcal{S}_1} - \xi_{m,0}^{\mathcal{S}_0}\|_{\mathcal{H}}^2) \mid \mathcal{Z}_{2m}] - \mathbb{E}[h(\|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2)] \right| \to 0 \quad (5.141)$$

almost surely. Now, we argue pointwise for each realization $(z_{2m})_{m\in\mathbb{N}}$ of $(\mathcal{Z}_{2m})_{m\in\mathbb{N}}$ such that (5.141) holds. As convergence in distribution implies convergence of CDF's at continuity points (Dudley, 2002, Theorem 9.3.6), this implies that

$$\mathbb{P}\left(\|\xi_{n,1}^{\mathcal{S}_1} - \xi_{n,0}^{\mathcal{S}_0}\|_{\mathcal{H}}^2 > z \middle| \, \mathcal{Z}_{2m} \right) \to \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2 > z)$$

almost surely for all $z \geq 0$.

Claim: There exists a further subsequence $l = n(\ell(m(l)))$ such that $\lim_l c_{l,\alpha} = c_\alpha$ exists that satisfies

$$\alpha \geq \mathbb{P}\left(\left\| \xi_{l,1}^{\mathcal{S}_1} - \xi_{l,0}^{\mathcal{S}_0} \right\|_{\mathcal{H}}^2 > c_{l,\alpha} \mid \mathcal{Z}_{2l} \right) \to \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|_{\mathcal{H}}^2 > c_\alpha) \quad (5.142)$$

almost surely. Moreover, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$, we also have

$$\mathbb{P}\left(\frac{1}{\sigma^2_{l,1}}\,\|\hat{\mu}_{l,1}(\mathbf{x}) - \hat{\mu}_{l,0}(\mathbf{x})\|^2_{\mathcal{H}} > c_{l,\alpha}\right) \to \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|^2_{\mathcal{H}} > c_\alpha). \quad (5.143)$$

Proof: First, since $c_{m,\alpha}$, $m \in \mathbb{N}$, is a bounded sequence as discussed above, we can find a convergent subsequence indexed by l such that $c_{l,\alpha} \to c_\alpha$, where $c_\alpha \in \mathbb{R}$ might depend on the chosen subsequence. Using Slutsky's theorem, we have that

$$\frac{1}{\sigma^2_{l,1}}\,\|\hat{\mu}_{l,1}(\mathbf{x}) - \hat{\mu}_{l,0}(\mathbf{x})\|^2_{\mathcal{H}} - c_{l,\alpha} \xrightarrow{D} \|\xi_1 - c_2(\mathbf{x})\xi_0\|^2_{\mathcal{H}} - c_\alpha.$$

Consequently, if $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$, we have

$$\mathbb{P}\left(\frac{1}{\sigma^2_{l,1}}\,\|\hat{\mu}_{l,1}(\mathbf{x}) - \hat{\mu}_{l,0}(\mathbf{x})\|^2_{\mathcal{H}} - c_{l,\alpha} > 0\right) \to \mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|^2_{\mathcal{H}} - c_\alpha > 0).$$

Similarly, arguing again pointwise for a realization z_l, $l \in \mathbb{N}$, and using Slutsky's theorem, we have

$$\mathbb{P}\left(\left\|\xi^{\mathcal{S}_1}_{l,1} - \xi^{\mathcal{S}_0}_{l,0}\right\|^2_{\mathcal{H}} - c_{l,\alpha} > 0 \mid \mathcal{Z}_{2l}\right) \to \mathbb{P}\left(\|\xi_1 - c_2(\mathbf{x})\xi_0\|^2_{\mathcal{H}} - c_\alpha > 0\right)$$

almost surely. □

We note that c_α, and thus the limit $\mathbb{P}(\|\xi_1 - c_2(\mathbf{x})\xi_0\|^2_{\mathcal{H}} > c_\alpha)$, might depend on the chosen subsequence. However, the α-bound in (5.142) holds by construction. Consequently, it follows from (5.143) that we have

$$\begin{aligned}\lim_l \mathbb{P}\left(\frac{1}{\sigma^2_{l,1}}\,\|\hat{\mu}_{l,1}(\mathbf{x}) - \hat{\mu}_{l,0}(\mathbf{x})\|^2_{\mathcal{H}} > c_{l,\alpha}\right) &= \lim_l \mathbb{P}\left(\left\|\xi^{\mathcal{S}_1}_{l,1} - \xi^{\mathcal{S}_0}_{l,0}\right\|^2_{\mathcal{H}} > c_{l,\alpha} \mid \mathcal{Z}_l\right) \\ &\leq \alpha \end{aligned} \quad (5.144)$$

almost surely under $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} = \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$.

Thus, we found that for every subsequence, there exists a further subsequence such that (5.144) holds. Now, assume that for the overall sequence

$$\limsup_n \mathbb{P}\left(\frac{1}{\sigma^2_{n,1}}\,\|\hat{\mu}_{n,0}(\mathbf{x}) - \hat{\mu}_{n,1}(\mathbf{x})\|^2_{\mathcal{H}} > c_{n,\alpha}\right) > \alpha.$$

Then, we can choose a subsequence satisfying

$$\lim_{m} \mathbb{P}\left(\frac{1}{\sigma_{m,1}^2}\left\|\hat{\mu}_{m,0}(\mathbf{x}) - \hat{\mu}_{m,1}(\mathbf{x})\right\|_{\mathcal{H}}^2 > c_{m,\alpha}\right) > \alpha.$$

But for this sequence, it is not possible to find a further subsequence l such that

$$\lim_{\ell} \mathbb{P}\left(\frac{1}{\sigma_{l,1}^2}\left\|\hat{\mu}_{l,0}(\mathbf{x}) - \hat{\mu}_{l,1}(\mathbf{x})\right\|_{\mathcal{H}}^2 > c_{l,\alpha}\right) \leq \alpha,$$

a contradiction to (5.144).

On the other hand, since (**K3**) holds, $\mathbb{P}^0_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}} \neq \mathbb{P}^1_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$ implies $\mu_1(\mathbf{x}) \neq \mu_2(\mathbf{x})$. Moreover, we have

$$\begin{aligned}
&\frac{1}{\sigma_{n,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \hat{\mu}_{n_0,0}(\mathbf{x})) \\
=&\frac{1}{\sigma_{n,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) + \frac{1}{\sigma_{n,1}}(\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})) - \frac{\sigma_{n,0}}{\sigma_{n,1}}\frac{1}{\sigma_{n,0}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x})).
\end{aligned}$$

Define

$$\xi_n^{01} = \frac{1}{\sigma_{n,1}}(\hat{\mu}_{n_1,1}(\mathbf{x}) - \mu_1(\mathbf{x})) - \frac{\sigma_{n,0}}{\sigma_{n,1}}\frac{1}{\sigma_{n,0}}(\hat{\mu}_{n_0,0}(\mathbf{x}) - \mu_0(\mathbf{x}))$$

Next, we have

$$\begin{aligned}
&\frac{1}{\sigma_{n,1}^2}\left\|\hat{\mu}_{n,1}(\mathbf{x}) - \hat{\mu}_{n,0}(\mathbf{x})\right\|_{\mathcal{H}}^2 \\
=&\|\xi_n^{01}\|_{\mathcal{H}}^2 + \frac{1}{\sigma_{n,1}^2}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}^2 + 2\left\langle \xi_n^{01}, \frac{1}{\sigma_{n,1}}(\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}))\right\rangle_{\mathcal{H}} \\
\geq&\|\xi_n^{01}\|_{\mathcal{H}}^2 + \frac{1}{\sigma_{n,1}^2}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}^2 - 2\left|\left\langle \xi_n^{01}, \frac{1}{\sigma_{n,1}}(\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}))\right\rangle_{\mathcal{H}}\right| \\
\geq&\|\xi_n^{01}\|_{\mathcal{H}}^2 + \frac{1}{\sigma_{n,1}^2}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}^2 - 2\|\xi_n^{01}\|_{\mathcal{H}}\frac{1}{\sigma_{n,1}}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}.
\end{aligned}$$

Due to $\|\xi_n^{01}\|_{\mathcal{H}}^2 \xrightarrow{D} \|\xi_0 - c_2(\mathbf{x})\xi_1\|_{\mathcal{H}}^2$, we infer

$$\|\xi_n^{01}\|_{\mathcal{H}}^2 = \mathcal{O}_p(1).$$

For the remaining terms, we have

$$\frac{1}{\sigma_{n,1}^2}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}^2 - 2\|\xi_n^{01}\|_{\mathcal{H}}\frac{1}{\sigma_{n,1}}\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}}$$

$$
\begin{aligned}
&= \frac{1}{\sigma_{n,1}} \|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}} \left(\frac{1}{\sigma_{n,1}} \|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}} - 2\|\xi_n^{01}\|_{\mathcal{H}} \right) \\
&\to \infty,
\end{aligned}
$$

as $\|\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})\|_{\mathcal{H}} > 0$ and $\sigma_{n,1} \to 0$. But since $\sup_n c_{n,\alpha} \leq M_\alpha$, this immediately implies that (ii) must hold.

□

Theorem 5.4.3. *Assume conditions (**F1**)–(**F5**) and (**D1**)–(**D7**) for the control and the treatment group, and assume that (**K1**) and (**K2**) hold together with strong ignorability and* (5.30)*. Then, for $\mathcal{B}(\mathbf{y})$ as in* (5.34)*, with $n_0, n_1 \to \infty$ such that $n_0/n_1 \to 1$,*

$$
\liminf_{n_0, n_1 \to \infty} \mathbb{P}\left(\cap_{\mathbf{y}} \{\mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}) \in \mathcal{B}(\mathbf{y})\} \right) \geq 1 - \alpha. \tag{5.35}
$$

Proof. Due to (**K2**), the kernel k is bounded. Without loss of generality, we assume that $C = 1$ bounds the kernel, so that $\sup_{\mathbf{y}} k(\mathbf{y}, \mathbf{y}) \leq 1$. Moreover, we assume again $n_0 = n_1 = n$, which does not affect asymptotics.

First, by definition of $\mathcal{B}(\mathbf{y})$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\forall \mathbf{y} \;\; \mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}) \in \mathcal{B}(\mathbf{y}) \right) \\
=&\mathbb{P}\left(\forall \mathbf{y} \;\; |\hat{\mu}_{n,1}(\mathbf{x})(\mathbf{y}) - \hat{\mu}_{n,0}(\mathbf{x})(\mathbf{y}) - (\mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}))| \leq \sqrt{c_{n,\alpha}} \sigma_{n,1} \right).
\end{aligned}
$$

The probability of the complementary event is given by

$$
\begin{aligned}
&\mathbb{P}\left(\exists \mathbf{y} \;\; |\hat{\mu}_{n,1}(\mathbf{x})(\mathbf{y}) - \hat{\mu}_{n,0}(\mathbf{x})(\mathbf{y}) - (\mu_1(\mathbf{x})(\mathbf{y}) - \mu_0(\mathbf{x})(\mathbf{y}))| > \sqrt{c_{n,\alpha}} \sigma_{n,1} \right) \\
=&\mathbb{P}\left(\exists \mathbf{y} \;\; |\langle \hat{\mu}_{n,1}(\mathbf{x}) - \hat{\mu}_{n,0}(\mathbf{x}) - (\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})), k(\mathbf{y}, \cdot) \rangle_{\mathcal{H}}| > \sqrt{c_{n,\alpha}} \sigma_{n,1} \right) \\
\leq&\mathbb{P}\left(\frac{1}{\sigma_{n,1}} \|\hat{\mu}_{n,1}(\mathbf{x}) - \hat{\mu}_{n,0}(\mathbf{x}) - (\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}))\|_{\mathcal{H}} > \sqrt{c_{n,\alpha}} \right)
\end{aligned}
$$

due to $\|k(\mathbf{y}, \cdot)\|_{\mathcal{H}}^2 = k(\mathbf{y}, \mathbf{y}) \leq 1$. By the same arguments as in the proof of Theorem 5.4.2 together with (5.27), we have

$$
\limsup_n \mathbb{P}\left(\frac{1}{\sigma_{n,1}^2} \|\hat{\mu}_{n,1}(\mathbf{x}) - \hat{\mu}_{n,0}(\mathbf{x}) - (\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}))\|_{\mathcal{H}}^2 > c_{n,\alpha} \right) \leq \alpha,
$$

implying (5.35). □

6 | The R-package `dmlalg`

The `dmlalg` package contains implementations of double machine learning algorithms in R.

6.1 | Installation

The released version of `dmlalg` can be installed from "The Comprehensive R Archive Network" (CRAN, `https://cran.r-project.org`) with the following command.

```
install.packages("dmlalg")
```

The package contains two sets of functions, one to estimate and make inference for linear parameters in a partially linear mixed-effects model for repeated measurements, and another one to estimate and make inference for linear parameters in a partially linear endogenous model using our regularization method. Subsequently, we detail these two sets of functions.

6.2 | Partially Linear Mixed-Effects Models for Repeated Measurements

The aim of this first set of functions is to estimate and perform inference for the linear coefficient in a partially linear mixed-effects model with DML. Machine learning algorithms allows us to incorporate more complex interaction structures and high-dimensional variables. This algorithm is described in Emmenegger and Bühlmann (2021a) and implemented in the function `mmdml`. This first set of functions consists of the following:

- `mmdml` computes the estimate of the linear parameter in a partially linear mixed-effects model using double machine learning methods.
- `confint` method for objects fitted with `mmdml`.
- `fixef` method for objects fitted with `mmdml`.
- `print` method for objects fitted with `mmdml`.
- `ranef` method for objects fitted with `mmdml`.
- `residuals` method for objects fitted with `mmdml`.
- `sigma` method for objects fitted with `mmdml`.
- `summary` method for objects fitted with `mmdml`.
- `vcov` method for objects fitted with `mmdml`.
- `VarCorr` method for objects fitted with `mmdml`.

6.2.1 | Example

This is a basic example which shows you how to solve a common problem:

```
library(dmlalg)

## generate data
RNGkind("L'Ecuyer-CMRG")
set.seed(19)
data1 ← example_data_mmdml(beta0 = 0.2)
data2 ← example_data_mmdml(beta0 = c(0.2, 0.2))

## fit models
## Caveat: Warning messages are displayed because the small number of
## observations results in a singular random effects model
fit1 ←
  mmdml(w = c("w1", "w2", "w3"), x = "x1", y = "resp", z = c("id", "
      cask"),
         data = data1, z_formula = "(1|id) + (1|cask:id)", group = "id",
             S = 3)
#> Warning in mmdml(w = c("w1", "w2", "w3"), x = "x1", y = "resp", z =
    c("id", :
#> Warning messages:
#> boundary (singular) fit: see ?isSingular

fit2 ←
  mmdml(w = c("w1", "w2", "w3"), x = c("x1", "x2"), y = "resp", z = c("
      id", "cask"),
         data = data2, z_formula = "(1|id) + (1|cask:id)", group = "id",
             S = 3)
#> Warning in mmdml(w = c("w1", "w2", "w3"), x = c("x1", "x2"), y = "
    resp", :
#> Warning messages:
#> boundary (singular) fit: see ?isSingular

## apply methods
confint(fit2)
#>           2.5%     97.5%
#> x1 -0.03415795 0.3480103
#> x2  0.15930098 0.3893938
fixef(fit2)
#>        x1        x2
#> 0.1569261 0.2743474
print(fit2)
#> Semiparametric mixed model fit by maximum likelihood ['mmdml']
#> Random effects:
#>  Groups   Name        Std.Dev.
#>  cask:id  (Intercept) 1.908e-06
#>  id       (Intercept) 1.107e-01
#>  Residual             2.756e-01
#> Number of obs: 46, groups:  cask:id, 20; id, 10
#> Fixed Effects:
#>     x1      x2
#> 0.1569  0.2743
#> optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer
    warnings; 1 lme4 warnings
ranef(fit2)
#> $`cask:id`
#>        (Intercept)
#> 1:1  -0.0023043914
#> 1:10 -0.0050894736
#> 1:2   0.0024571669
#> 1:3   0.0007708872
#> 1:4  -0.0012417525
```

```
#> 1:5    0.0029010344
#> 1:6    0.0012307712
#> 1:7   -0.0028418387
#> 1:8   -0.0015618712
#> 1:9   -0.0048037635
#> 2:1    0.0100768089
#> 2:10 -0.0031560819
#> 2:2   -0.0033427429
#> 2:3   -0.0044928425
#> 2:4   -0.0054049237
#> 2:5   -0.0021157461
#> 2:6   -0.0023122280
#> 2:7    0.0038004751
#> 2:8    0.0148222090
#> 2:9    0.0026385335
#>
#> $id
#>     (Intercept)
#> 1    0.100740957
#> 10 -0.124434023
#> 2   -0.036918731
#> 3   -0.030230821
#> 4   -0.081051109
#> 5    0.018887512
#> 6   -0.006711504
#> 7    0.025545300
#> 8    0.235373382
#> 9   -0.020965920
residuals(fit2)
#> [[1]]
#>  [1] -0.1311195998  0.5733692328  0.1398125051 -0.0705463911
    -0.1196552839
#>  [6] -0.0354080600  0.6205378654 -0.1057642425 -0.4355021749
    -0.0633888854
#> [11]  0.0070044016 -0.1777683530 -0.0214893719  0.0052358066
    0.1594839987
#> [16] -0.2353753755 -0.2216497409 -0.1034882421  0.0175984650
    -0.0388497525
#> [21]  0.4636325671 -0.2597143034  0.3528825573 -0.4739722035
    0.0007039458
#> [26]  0.0700307380 -0.1315655000 -0.1617002846  0.2162465843
    0.0934414339
#> [31] -0.0480554546 -0.1342562672 -0.2349311153  0.4021334289
    0.6761796261
#> [36]  0.3514207835 -0.0918140917 -0.2144924370 -0.3184478283
    -0.2704273590
#> [41] -0.1953366308  0.7209607369 -0.1050645053 -0.2895904461
    -0.2737160112
#> [46]  0.0941353224
#>
#> [[2]]
#>  [1]  0.066708484  0.381936532  0.083961541 -0.244607521
    -0.116940987
#>  [6] -0.015024540  0.605540877  0.128223071 -0.186010749
    -0.119432458
#> [11] -0.101885530 -0.153724682 -0.214346785 -0.126400135
    0.090522034
#> [16] -0.103818112 -0.170763502 -0.102507199  0.047067741
    -0.026325741
#> [21]  0.472126666 -0.231575911  0.324749223 -0.423215690
    -0.013990681
```

```
#> [26]  0.066537726 -0.086954711 -0.025470109  0.227756255
    0.224213587
#> [31] -0.070700603  0.081484834 -0.226268534  0.615553468
    0.723110460
#> [36]  0.333538915 -0.076459138 -0.198241935 -0.245660371
    -0.366166157
#> [41] -0.142947352  0.677671159 -0.047532882 -0.305555800
    -0.379445954
#> [46]  0.007159723
#>
#> [[3]]
#>  [1]  0.09685066  0.34629638  0.09582384 -0.27150981 -0.12653048
    -0.02387543
#>  [7]  0.62488259  0.12730531 -0.19466784 -0.12227940 -0.07635676
    -0.16470188
#> [13] -0.20223445 -0.11432450  0.13844295 -0.12234863 -0.18662475
    -0.09034621
#> [19]  0.07330126 -0.02704395  0.51049151 -0.23716208  0.36116367
    -0.42669942
#> [25] -0.02948530  0.10139429 -0.06858354 -0.03611104  0.19153360
    0.21971922
#> [31] -0.04085530  0.09453877 -0.20903814  0.60734696  0.69658489
    0.33318587
#> [37] -0.09082740 -0.21317885 -0.24276713 -0.34992920 -0.09491974
    0.68198892
#> [43] -0.07291051 -0.24350682 -0.40714805  0.05067157
sigma(fit2)
#> [1] 0.2756384
summary(fit2)
#> Semiparametric mixed model fit by maximum likelihood ['mmdml']
#> Scaled residuals (nr_res = 3):
#>     Min      1Q  Median      3Q     Max
#> -1.7195 -0.6674 -0.2394  0.3637  2.6234
#>
#> Random effects:
#>  Groups   Name        Variance  Std.Dev.
#>  cask:id  (Intercept) 3.641e-12 1.908e-06
#>  id       (Intercept) 1.226e-02 1.107e-01
#>  Residual             7.598e-02 2.756e-01
#> Number of obs: 46, groups:  cask:id, 20; id, 10
#>
#> Fixed effects:
#>    Estimate Std. Error z value Pr(>|z|)
#> x1  0.15693    0.09749   1.610    0.107
#> x2  0.27435    0.05870   4.674 2.96e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>    x1
#> x2 -0.029
#> optimizer (nloptwrap) convergence code: 0 (OK)
#> boundary (singular) fit: see ?isSingular
vcov(fit2)
#> 2 x 2 Matrix of class "dpoMatrix"
#>               x1            x2
#> x1  9.505018e-03 -9.208662e-05
#> x2 -9.208662e-05  3.445483e-03
VarCorr(fit2)
#>  Groups   Name        Std.Dev.
#>  cask:id  (Intercept) 1.9083e-06
```

```
#>  id        (Intercept) 1.1074e-01
#>  Residual              2.7564e-01
```

6.3 | Partially linear models with confounding variables

The aim of this second set of functions it to perform inference for the linear parameter in partially linear models with confounding variables. The standard DML estimator of the linear parameter has a two-stage least squares interpretation, which can lead to a large variance and overwide confidence intervals. We apply regularization to reduce the variance of the estimator, which produces narrower confidence intervals that remain approximately valid. Nuisance terms can be flexibly estimated with machine learning algorithms. This algorithm is described in Emmenegger and Bühlmann (2021) and implemented in the function `regsdml`. This second set of functions consists of the following:

- `regsdml` computes the estimate of the linear parameter in a partially linear model with endogenous variables with regularized and standard double machine learning methods.
- `summary` method for objects fitted with `regsdml`.
- `confint` method for objects fitted with `regsdml`.
- `coef` method for objects fitted with `regsdml`.
- `vcov` method for objects fitted with `regsdml`.
- `print` method for objects fitted with `regsdml`.

6.3.1 | Example

This is a basic example which shows you how to solve a common problem:

```
library(dmlalg)

## Generate some data:
set.seed(19)
# true linear parameter
beta0 ← 1
n ← 40
# observed confounder
w ← pi * runif(n, -1, 1)
# instrument
a ← 3 * tanh(2 * w) + rnorm(n, 0, 1)
# unobserved confounder
h ← 2 * sin(w) + rnorm(n, 0, 1)
# linear covariate
x ← -1 * abs(a) - h - 2 * tanh(w) + rnorm(n, 0, 1)
# response
y ← beta0 * x - 3 * cos(pi * 0.25 * h) + 0.5 * w ^ 2 + rnorm(n, 0, 1)

## Estimate the linear coefficient from x to y
## (The parameters are chosen small enough to make estimation fast):
## Caveat: A spline estimator is extrapolated, which raises a warning
    message.
```

```
## Extrapolation lies in the nature of our method. To omit the warning
    message
## resulting from the spline estimator, another estimator may be used.
fit ← regsdml(a, w, x, y,
                gamma = exp(seq(-4, 1, length.out = 4)),
                S = 3,
                do_regDML_all_gamma = TRUE,
                cond_method = c("forest",  # for E[A|W]
                                "spline",  # for E[X|W]
                                "spline"), # for E[Y|W]
                params = list(list(ntree = 1), NULL, NULL))
#> Warning in print_W_E_fun(errors, warningMsgs):
#> Warning messages:
#> some 'x' values beyond boundary knots may cause ill-conditioned
    bases
## parm = c(2, 3) prints an additional summary for the 2nd and 3rd
    gamma-values
summary(fit, parm = c(2, 3),
        correlation = TRUE,
        print_gamma = TRUE)
#>
#> Coefficients :
#> regsDML (2.72e+00) :
#>    Estimate Std. Error  z value     Pr(>|z|)
#> b1 0.910255  0.1731559 5.256852 1.465421e-07
#>
#> regDMLall (9.70e-02) :
#>     Estimate Std. Error  z value     Pr(>|z|)
#> b1 0.7986392  0.1514027 5.274935 1.328031e-07
#>
#> regDMLall (5.13e-01) :
#>    Estimate Std. Error  z value     Pr(>|z|)
#> b1 0.846176  0.1651298 5.124308 2.986318e-07
#>
#>
#> Variance-covariance matrices :
#> regsDML (2.72e+00) :
#>             b1
#> b1 0.02998297
#>
#> regDMLall (9.70e-02) :
#>             b1
#> b1 0.02292277
#>
#> regDMLall (5.13e-01) :
#>             b1
#> b1 0.02726785
confint(fit, parm = c(2, 3),
        print_gamma = TRUE)
#>
#> Two-sided confidence intervals at level 0.95 :
#>
#> regsDML (2.72e+00) :
#>        2.5 %   97.5 %
#> b1 0.5708757 1.249634
#>
#> regDMLall (9.70e-02) :
#>        2.5 %   97.5 %
#> b1 0.5018955 1.095383
#>
#> regDMLall (5.13e-01) :
```

```
#>        2.5 %   97.5 %
#> b1 0.5225276 1.169824
coef(fit) # coefficients
#>     regsDML
#> b1 0.910255
vcov(fit) # variance-covariance matrices
#>
#> Variance-covariance matrices :
#> regsDML :
#>             b1
#> b1 0.02998297

## Alternatively, provide the data in a single data frame
## (see also caveat above):
data ← data.frame(a = a, w = w, x = x, y = y)
fit ← regsdml(a = "a", w = "w", x = "x", y = "y", data = data,
              gamma = exp(seq(-4, 1, length.out = 4)),
              S = 3)
#> Warning in print_W_E_fun(errors, warningMsgs):
#> Warning messages:
#> some 'x' values beyond boundary knots may cause ill-conditioned
    bases

## With more realistic parameter choices:
if (FALSE) {
  fit ← regsdml(a, w, x, y,
                cond_method = c("forest",  # for E[A|W]
                                "spline",  # for E[X|W]
                                "spline")) # for E[Y|W]
  summary(fit)
  confint(fit)

  ## Alternatively, provide the data in a single data frame:
  ## (see also caveat above):
  data ← data.frame(a = a, w = w, x = x, y = y)
  fit ← regsdml(a = "a", w = "w", x = "x", y = "y", data = data)
}
```

Bibliography

D. Acemoglu, S. Johnson, and J. A. Robinson. The colonial origins of comparative development: An empirical investigation. The American Economic Review, 91(5):1369–1401, 2001.

K. Aggarwal, M. Kirchmeyer, P. Yadav, S. S. Keerthi, and P. Gallinari. Benchmarking regression methods: A comparison with CGAN. Preprint arXiv:1905.12868, 2019.

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica, 71(6):1795–1843, 2003.

A. M. Aiken, C. Davey, J. R. Hargreaves, and R. J. Hayes. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication. International Journal of Epidemiology, 44(6):1572–1580, 2015.

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

T. Amemiya. Advanced Econometrics. Harvard University Press, Cambridge, Massachusetts, 1985.

T. Anderson, N. Kunitomo, and Y. Matsushita. On the asymptotic optimality of the LIML estimator with possibly many instruments. Journal of Econometrics, 157(2):191–204, 2010.

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

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

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. The Annals of Mathematical Statistics, 20(1):46–63, 1949.

T. W. Anderson and T. Sawa. Evaluation of the distribution function of the two-stage least squares estimate. Econometrica, 47(1):163–182, 1979.

T. W. Anderson, N. Kunitomo, and T. Sawa. Evaluation of the distribution function of the limited information maximum likelihood estimator. Econometrica, 50(4):1009–1027, 1982.

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

D. W. Andrews. Empirical process methods in econometrics. In R. F. Engle and D. McFadden, editors, Handbook of Econometrics, volume 4 of Handbook of Econometrics, chapter 37, pages 2247–2294. Elsevier, 1986.

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

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444–455, 1996.

T. T. Aniley, L. K. Debusho, Z. M. Nigusie, W. K. Yimer, and B. B. Yimer. A semi-parametric mixed models for longitudinally measured fasting blood sugar level of adult diabetic patients. BMC Medical Research Methodology, 19(13), 2019.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. The Annals of Statistics, 47(2):1148–1178, 2019.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–972, 2005.

R. L. Basmann. A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica, 25(1):77–83, 1957.

A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1):82–130, 2011.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2):521–547, 2013.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika, 98(4):791–806, 2011.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6):2369–2429, 2012.

E. R. Berndt, B. H. Hall, R. E. Hall, and J. A. Hausman. Estimation and inference in nonlinear structural models. Annals of Economic and Social Measurement, 3(4):653–665, 1974.

P. J. Bickel. On adaptive estimation. The Annals of Statistics, 10(3):647–671, 1982.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics, 66(4): 1069–1077, 2010.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. The Annals of Probability, 33(2):514–560, 2005.

J. Bound, D. A. Jaeger, and R. M. Baker. 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(430):443–450, 1995.

R. J. Bowden and D. A. Turkington. Instrumental variables. Econometric Society Monographs. Cambridge University Press, Cambridge, 1985.

J. Bradic, G. Claeskens, and T. Gueuning. Fixed effects testing in high-dimensional linear mixed models. Journal of the American Statistical Association, 115(532):1835–1850, 2020.

L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

P. Bühlmann. Invariance, causality and robustness. Statistical Science, 35(3): 404–426, 2020.

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

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

J. Cai, A. De Janvry, and E. Sadoulet. Social networks and the decision to insure. American Economic Journal: Applied Economics, 7(2):81–108, 2015.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 35(6):2313–2351, 2007.

D. Carl, C. Emmenegger, W. Yuan, M. Zheng, and Z. Guo. TSCI: Tools for causal inference with possibly invalid instrumental variables, 2022. URL `https://cran.r-project.org/web/packages/TSCI/index.html`. R-package available on CRAN.

D. Ćevid, L. Michel, J. Näf, N. Meinshausen, and P. Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. Journal of Machine Learning Research, 23(333):1–79, 2022.

T. Chamorro-Premuzic and A. Furnham. Personality, intelligence and approaches to learning as predictors of academic performance. Personality and Individual Differences, 44(7):1596–1603, 2008.

B. Chen, H. Liang, and Y. Zhou. GMM estimation in partial linear models with endogenous covariates causing an over-identified problem. Communications in Statistics - Theory and Methods, 45(11):3168–3184, 2016.

J. Chen, C.-H. Huang, and J.-J. Tien. Debiased/double machine learning for instrumental variable quantile regressions. Econometrics, 9(2), 2021.

L. Chen and H. Cao. Analysis of asynchronous longitudinal data with partially linear models. Electronic Journal of Statistics, 11(1):1549–1569, 2017.

X. Chen and H. White. Central limit and functional central limit theorems for hilbert-valued dependent heterogeneous arrays with applications. Econometric Theory, 14(2):260–284, 1998.

X. Chen and H. White. Improved rates and asymptotic normality for nonparametric neural network estimators. IEEE Transactions on Information Theory, 45:682–691, 1999.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. The Annals of Statistics, 42(4):1564–1597, 2014.

V. Chernozhukov, C. Hansen, and M. Spindler. hdm: High-dimensional metrics. R Journal, 8(2):185–199, 2016.

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

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 2018.

U. Cherubini, E. Luciano, and W. Vecchiato. Copula methods in finance. John Wiley & Sons, 2004.

H. D. Chiang, K. Kato, Y. Ma, and Y. Sasaki. Multiway cluster robust double/debiased machine learning. Journal of Business & Economic Statistics, 0(0):1–11, 2021.

A. Chin. Central limit theorems via stein's method for randomized experiments under interference, 2018. Preprint arXiv:1804.03105.

A. Chin. Regression adjustments for estimating the global treatment effect in experiments with interference. Journal of Causal Inference, 7(2), 2019.

S. Cohen and G. Williamson. Perceived stress in a probability sample of the United States. In S. Spacapam and S. Oskamp, editors, The Social Psychology of Health: Claremont Symposium on Applied Social Psychology. Sage, Newbury Park, CA, 1988.

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

P. Constantinou and A. P. Dawid. Extended conditional independence and applications in causal inference. The Annals of Statistics, 45(6):2618–2653, 2017.

J. G. Cragg. On the relative small-sample properties of several structural-equation estimators. Econometrica, 35(1):89–110, 1967.

W. H. Crown, H. J. Henk, and D. J. Vanness. 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(8):1078–1084, 2011.

G. Csardi and T. Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006. URL `https://igraph.org`.

Y. Cui and E. Tchetgen Tchetgen. Selective machine learning of doubly robust functionals, 2020. Preprint arXiv:1911.02029.

G. Daraganova and G. Robins. Autologistic Actor Attribute Models, pages 102–114. Structural Analysis in the Social Sciences. Cambridge University Press, 2012.

A. DasGupta. Asymptotic theory of statistics and probability. Springer Texts in Statistics. Springer, New York, 2008.

M. Davidian and D. M. Giltinan. Nonlinear models for repeated measurement data, volume 62 of Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, Florida, 1995.

M. Davidian and D. M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. Journal of Agricultural, Biological, and Environmental Statistics, 8(4):387–419, 2003.

C. S. Davis. Statistical methods for the analysis of repeated measurements. Springer Texts in Statistics. Springer, New York, 2002.

E. Demidenko. Mixed Models: Theory and Applications. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 2004.

K. DiazOrdaz, R. Daniel, and N. Kreif. Data-adaptive doubly robust instrumental variable methods for treatment effect heterogeneity, 2019. Preprint arXiv:1802.02821.

R. M. Dudley. Real Analysis and Probability. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002.

R. Durrett. Probability: theory and examples. Duxbury Press, Belmont, CA, fourth edition, 1996.

D. Eckles and E. Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. Journal of the American Statistical Association, 116 (534):507–517, 2021.

D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. Journal of Causal Inference, 5 (1):20150021, 2017.

N. Egami and E. J. Tchetgen Tchetgen. Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding, 2021. Preprint arXiv:2109.01933.

C. Emmenegger. dmlalg: Double machine learning algorithms, 2021. URL `https://cran.r-project.org/web/packages/dmlalg/index.html`. R-package available on CRAN.

C. Emmenegger. Double machine learning methods: Beyond independence. In Re-thinking High-dimensional Mathematical Statistics (Workshop Report), volume 25, pages 21–23. Oberwolfach Reports, Oberwolfach, Germany, 5 2022a. URL `http://publications.mfo.de/handle/mfo/3962`. Organizers: Florentina Bunea, Robert Nowak, Alexandre Tsybakov.

C. Emmenegger. Analyzing longitudinal data with plug-in machine learning. Bulletin of the Swiss Statistical Society, 101:8, 4 2022b. URL `https://www.stat.ch/images/bulletin/pdfs/bulletin_101.pdf`. Poster presented at the Swiss Statistical Seminar 2022, March 25.

C. Emmenegger. Regularizing double machine learning in partially linear endogenous models. Copenhagen Causality Lab (CoCaLa), Department of Mathematical Sciences, University of Copenhagen, April 2021.

C. Emmenegger. Double machine learning in partially linear models. Bernoulli's Round table, Department of Mathematics and Computer Science, University of Basel, December 2021.

C. Emmenegger. Regularized double machine learning in partially linear models with unobserved confounding. Bernoulli-IMS 10th World Congress in Probability and Statistics, July 2021.

C. Emmenegger. Regularized double machine learning in partially linear models with unobserved confounding. ISI 63rd World Statistics Congress, June 2021.

C. Emmenegger. Analyzing longitudinal data with plug-in machine learning. Swiss Statistics Seminar, March 2022.

C. Emmenegger. Double machine learning methods: Beyond independence. Oberwolfach Workshop "Re-thinking High-dimensional Mathematical Statistics", May 2022a.

C. Emmenegger. Machine learning for repeated measurements data. Women in Data Science, May 2022b.

C. Emmenegger. Mehr Glück im Spiel dank Statistik (engl.: more luck in game thanks to statistics). Känguru Schweiz, Kangaroo goes Science day, spring 2013.

C. Emmenegger and P. Bühlmann. Regularizing double machine learning in partially linear endogenous models. Electronic Journal of Statistics, 15(2): 6461–6543, 2021.

C. Emmenegger and P. Bühlmann. Plugin machine learning for partially linear mixed-effects models with repeated measurements, 2021a. Preprint arXiv:2108.13657.

C. Emmenegger and P. Bühlmann. Regularized double machine learning in partially linear models with unobserved confounding. In Proceedings of 63rd ISI World Statistics Congress, 2021b. URL `https://www.isi-web.org/files/docs/papers-and-abstracts/88-day2-cps020-regularized-double-machine-lea.pdf`.

C. Emmenegger, M.-L. Spohn, T. Elmer, and P. Bühlmann. Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning, 2022. Preprint arXiv:2206.14591.

P. Erdős and A. Rényi. On random graphs I. Publicationes Mathematicae, 6: 290–297, 1959.

L. Fahrmeir and T. Kneib. Bayesian smoothing and regression for longitudinal, spatial and event history data, volume 36 of Oxford statistical science series. Oxford University Press, New York, 2011.

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(2):303–322, 2000.

H. Farbmacher, M. Huber, L. Lafférs, H. Langen, and M. Spindler. Causal mediation analysis with double machine learning, 2020. Preprint arXiv:2002.12710.

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. Applied Longitudinal Analysis. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey, 2 edition, 2011.

J.-P. Florens, J. Johannes, and S. Van Bellegem. Instrumental regression in partially linear models. The Econometrics Journal, 15(2):304–324, 2012.

L. Forastiere, E. M. Airoldi, and F. Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. Journal of the American Statistical Association, 116(534):901–918, 2021.

W. A. Fuller. Some properties of a modification of the limited information estimator. Econometrica, 45(4):939–53, 1977.

W. A. Fuller. Measurement error models. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York, 1987.

G. González-Rodríguez and A. Colubi. On the consistency of bootstrap methods in separable hilbert spaces. Econometrics and Statistics, 1:118–127, 2017.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In Advances in neural information processing systems, pages 513–520, 2007.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In Advances in neural information processing systems, pages 1205–1213, 2012.

Q. Guoyou and Z. Zhongyi. Robust estimation in generalized semiparametric mixed models for longitudinal data. Journal of Multivariate Analysis, 98(8): 1658–1683, 2007.

Q. Guoyou and Z. Zhongyi. Robust estimation in partial linear mixed model for longitudinal data. Acta Mathematica Scientia, 28(2):333–347, 2008.

Q. Guoyou and Z. Zhongyi. Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data. Biometrics, 65 (1):52–59, 2009.

J. Hahn, J. Hausman, and G. Kuersteiner. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. The Econometrics Journal, 7(1):272–306, 2004.

J. Hájek. Comment on "an essay on the logical foundations of survey sampling, part one" by Basu. In V. P. Godambe and D. A. Sprott, editors, Foundations of Statistical Inference, page 236. Holt, Rinehart and Winston, Toronto, 1971.

E. Halloran and M. G. Hudgens. Dependent happenings: a recent methodological review. Current Epidemiology Reports, 3(4):297–305, 2016.

B. E. Hansen. Econometrics. University of Wisconsin, Department of Economics, 2017. Last revised on January 5, 2017.

L. P. Hansen. Large sample properties of generalized method of moments estimators. Econometrica, 50(4):1029–1054, 1982.

L. P. Hansen. A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. Journal of Econometrics, 30(1):203–238, 1985.

W. Härdle, H. Liang, and J. Gao. Partially linear models. Contributions to Statistics. Springer, Berlin Heidelberg, 2000.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. Nonparametric and semiparametric models. Springer series in statistics. Springer, Berlin, 2004.

J. D. Hart and T. E. Wehrly. Kernel regression estimation using repeated measurements data. Journal of the American Statistical Association, 81 (396):1080–1088, 1986.

J. J. Heckman. Skill formation and the economics of investing in disadvantaged children. Science, 312(5782):1900–1902, 2006.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. SIAM Review, 23(1):53–60, 1981.

R. C. Hill, W. E. Griffiths, and G. C. Lim. Principles of econometrics. Wiley, Hoboken, New Jersey, 4 edition, 2011.

G. H. Hillier and C. L. Skeels. Some further exact results for structural equation estimators. In P. C. B. Phillips, editor, Models, Methods and Applications of Econometrics: essays in Honor of A. R. Bergstroms, pages 117–139. Blackwell, Cambridge, Massachusetts, 1993.

K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

G. Hong and S. W. Raudenbush. Causal inference for time-varying instructional treatments. Journal of Educational and Behavioral Statistics, 33(3):333–362, 2008.

J. L. Horowitz. Applied nonparametric instrumental variables estimation. Econometrica, 79(2):347–394, 2011.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260):663–685, 1952.

T. Hoshino. Estimating a continuous treatment model with spillovers: A control function approach, 2021. Preprint arXiv:2112.15114.

T. Hsing and R. Eubank. Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics. Wiley, 2015.

Y. Hu, S. Li, and S. Wager. Average direct and indirect causal effects under interference. Biometrika, 02 2022.

M. G. Hudgens and E. Halloran. Toward causal inference with interference. Journal of the American Statistical Association, 103(482):832–842, 2008.

J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. Biometrics, 67(2):495–503, 2011.

M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE, 2020. Preprint arXiv:2005.03353.

S. Kim, D. Zeng, and J. M. G. Taylor. Joint partially linear model for longitudinal data with informative drop-outs. Biometrics, 73(1):72–82, 2017.

M. C. Knaus. Double machine learning based program evaluation under unconfoundedness, 2020. Preprint arXiv:2003.03191.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. The Annals of Statistics, 38(6):3660–3695, 2010.

M. R. Kosorok. Bootstraps of sums of independent but not identically distributed stochastic processes. Journal of Multivariate Analysis, 84(2):299–318, 2003.

M. R. Kosorok. Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics. Springer New York, 2008.

D. Kozbur. Analysis of testing-based forward model selection. Econometrica, 88(5):2147–2173, 2020.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10):4156–4165, 2019.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. Biometrics, 38(4):963–974, 1982.

T. Lattimore and C. Szepesvári. Bandit algorithms. Cambridge University Press, Cambridge, 2020.

S. L. Lauritzen. Graphical models. Oxford statistical science series. Clarendon Press, Oxford, 1996.

S. L. Lauritzen and T. S. Richardson. Chain graph models and their causal interpretations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):321–348, 2002.

Y. Lee and E. L. Ogburn. Network dependence can lead to spurious associations and invalid inference. Journal of the American Statistical Association, 116 (535):1060–1074, 2021.

M. Leung. Treatment and spillover effects under network interference. The Review of Economics and Statistics, 102(2):368–380, 2020.

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

S. Li and S. Wager. Random graph asymptotics for treatment effect estimation under network interference. The Annals of Statistics, 50(4):2334–2358, 2022.

S. Li, T. T. Cai, and H. Li. Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. Journal of the American Statistical Association, 2021.

Z. Li and L. Zhu. On variance components in semiparametric mixed models for longitudinal data. Scandinavian Journal of Statistics, 37(3):442–457, 2010.

H. Liang. Generalized partially linear mixed-effects models incorporating mismeasured covariates. Annals of the Institute of Statistical Mathematics, 61:27–46, 2009.

H. Lin, G. Qin, J. Zhang, and W. K. Fung. Doubly robust estimation of partially linear models for longitudinal data with dropouts and measurement error in covariates. Statistics, 52(1):84–98, 2018.

X. Lin and R. Carroll. Semiparametric regression for clustered data using generalized estimating equations. Journal of the American Statistical Association, 96(455):1045–1056, 2001.

L. Liu, M. G. Hudgens, B. Saul, J. D. Clemens, M. Ali, and M. E. Emch. Doubly robust estimation in observational studies with partial interference. Stat, 8(1):e214, 2019.

M. Liu, Y. Zhang, and D. Zhou. Double/debiased machine learning for logistic partially linear model. The Econometrics Journal, 2021.

W. P. Lloyd. A note on the use of the two-stage least squares estimator in financial models. The Journal of Financial and Quantitative Analysis, 10(1): 143–149, 1975.

R. Los and A. Schweinle. The interaction between student motivation and the instructional environment on academic outcome: a hierarchical linear model. Social Psychology of Education, 22(2):471–500, Apr. 2019.

T. Lu. Skew-t partially linear mixed-effects models for AIDS clinical studies. Journal of Biopharmaceutical Statistics, 26(5):899–911, 2016.

X. Luo, D. S. Small, C.-S. R. Li, and P. R. Rosenbaum. Inference with interference between units in an fmri experiment of motor inhibition. Journal of the American Statistical Association, 107(498):530–541, 2012.

Y. Luo and M. Spindler. High-dimensional $l_2$boosting: Rate of convergence, 2016. Preprint arXiv:1602.08927.

Y. Ma and R. J. Carroll. Locally efficient estimators for semiparametric models with measurement error. Journal of the American Statistical Association, 101(476):1465–1474, 2006.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors. Handbook of graphical models. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, Boca Raton, FL, 2019.

E. Mammen and S. van de Geer. Penalized quasi-likelihood estimation in partial linear models. The Annals of Statistics, 25(3):1014–1035, 1997.

C. F. Manski. Identification of endogenous social effects: The reflection problem. The Review of Economic Studies, 60(3):531–542, 07 1993.

R. S. Mariano. The existence of moments of the ordinary least squares and two-stage least squares estimators. Econometrica, 40(4):643–652, 1972.

R. S. Mariano. Analytical small-sample distribution theory in econometrics: The simultaneous-equations case. International Economic Review, 23(3): 503–533, 1982.

R. S. Mariano. Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications, chapter 6, pages 122–141. John Wiley & Sons, Ltd, 2003.

C. Masci, A. M. Paganoni, and F. Ieva. Semiparametric mixed effects models for unsupervised classification of Italian schools. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4):1313–1342, 2019.

L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. The Annals of Statistics, 37(6B):3779–3821, 2009.

N. Meinshausen. Quantile regression forests. Journal of Machine Learning Research, 7(Jun):983–999, 2006.

N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. Journal of the American Statistical Association, 104(488):1671–1681, 2009.

L. Michel and D. Ćevid. drf: Distributional Random Forests, 2021. URL `https://CRAN.R-project.org/package=drf`. R package version 1.1.0.

E. Miguel and M. Kremer. Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1):159–217, 2003.

H. Q. Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. Constructive Approximation, 32(2):307–338, Oct 2010.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends in Machine Learning, 10(1–2):1–141, 2017.

J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Inference for the distributional random forest, 2023. Preprint on arXiv:2302.05761.

A. L. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. Econometrica, 27(4):575–595, 1959.

A. L. Nagar. A monte carlo study of alternative simultaneous equation estimators. Econometrica, 28(3):573–590, 1960.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In Handbook of Econometrics, volume 4, chapter 36, pages 2111–2245. Elsevier Science, 1994.

E. L. Ogburn and T. J. VanderWeele. Causal diagrams for interference. Statistical science, 29(4):559–578, 2014.

E. L. Ogburn and T. J. VanderWeele. Vaccines, contagion, and social networks. The Annals of Applied Statistics, 11(2):919–948, 2017.

E. L. Ogburn, O. Sofrygin, I. Díaz, and M. J. van der Laan. Causal inference for social network data. Journal of the American Statistical Association, 0 (0):1–15, 2022.

R. Ohinata. Three Essays on Application of Semiparametric Regression: Partially Linear Mixed Effects Model and Index Model. PhD thesis, Wirtschaftswissenschaftlichen Fakultät der Universität Göttingen, Göttingen, Germany, 12 2012.

R. Okui, D. S. Small, Z. Tan, and J. M. Robins. Doubly robust instrumental variable regression. Statistica Sinica, 22(1):173–205, 2012.

J. Pan and Y. Pan. jmcm: An R package for joint mean-covariance modeling of longitudinal data. Journal of Statistical Software, 82(9):1–29, 2017.

G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In Proceedings of 38th International Conference on Machine Learning (ICML), volume 139 of Proceedings of Machine Learning Research, pages 8401–8412. PMLR, July 2021.

J. Pearl. Causal diagrams for empirical research. Biometrika, 82(4):669–688, 1995.

J. Pearl. Graphs, causality, and structural equation models. Sociological Methods & Research, 27(2):226–284, 1998.

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

J. Pearl. Causality: Models, reasoning, and inference. Cambridge University Press, Cambridge, 2 edition, 2009.

J. Pearl. An introduction to causal inference. The International Journal of Biostatistics, 6(2): Article 7, 2010.

C. Perez-Heydrich, M. G. Hudgens, E. Halloran, J. D. Clemens, M. Ali, and M. E. Emch. Assessing effects of cholera vaccination in the presence of interference. Biometrics, 70(3):731–741, 2014.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence

classes of ancestral graphs. Journal of Machine Learning Research, 18(220): 1–62, 2018.

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

K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012. URL `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`. Version 20121115.

P. C. B. Phillips. The exact distribution of LIML: I. International Economic Review, 25(1):249–261, 1984.

P. C. B. Phillips. The exact distribution of LIML: II. International Economic Review, 26(1):21–36, 1985.

J. C. Pinheiro. Topics in Mixed Effects Models. PhD thesis, University of Wisconsin, Madison, 1994.

J. C. Pinheiro and D. M. Bates. Mixed-effects models in S and S-PLUS. Statistics and computing. Springer, New York, 2000.

G. Pisier. Martingales in Banach Spaces. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

J. Praestgaard and J. A. Wellner. Exchangeably Weighted Bootstraps of the General Empirical Process. The Annals of Probability, 21(4):2053–2086, 1993.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society. Series B (Methodological), 53(1):233–243, 1991.

G. Robins, P. Pattison, and P. Elliott. Network models for social influence processes. Psychometrika, 66(2):161–189, 2001.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, 7(9):1393–1512, 1986.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(429):122–129, 1995.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 90(429):106–121, 1995.

P. M. Robinson. Root-N-consistent semiparametric regression. Econometrica, 56(4):931–954, 1988.

P. R. Rosenbaum. Model-based direct adjustment. Journal of the American Statistical Association, 82(398):387–394, 1987.

N. Ross. Fundamentals of Stein's method. Probability Surveys, 8(none):210–293, 2011.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

D. Rubin. Comment on: "randomization analysis of experimental data in the fisher randomization test" by D. Basu. Journal of the American Statistical Association, 75:591–593, 1980.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974.

D. Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric regression, volume 12 of Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, 2003.

F. Sävje, P. M. Aronow, and M. G. Hudgens. Average treatment effects in the presence of unknown interference. The Annals of Statistics, 49(2):673–701, 2021.

A. Scharfstein, Daniel O.and Rotnitzky and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120, 1999.

J. Schelldorfer, P. Bühlmann, and S. van de Geer. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. Scandinavian Journal of Statistics, 38(2):197–214, 2011.

C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research, 40(2):211–239, 2011.

C.-J. Simon-Gabriel, A. Barp, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. Preprint arXiv:2006.09268, 2020.

A. Sklar. Fonctions de Répartition À N Dimensions Et Leurs Marges. Université Paris 8, 1959.

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

T. A. Snijders, G. G. van de Bunt, and C. E. Steglich. Introduction to stochastic actor-based models for network dynamics. Social Networks, 32(1):44–60, 2010. Dynamics of Social Networks.

T. A. B. Snijders. Models for Longitudinal Network Data, pages 215–247. Structural Analysis in the Social Sciences. Cambridge University Press, 2005.

M. E. Sobel. What do randomized studies of housing mobility demonstrate? Journal of the American Statistical Association, 101(476):1398–1407, 2006.

O. Sofrygin and M. J. van der Laan. Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. Journal of Causal Inference, 5(1):1–35, 2017.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28:3483–3491, 2015.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 961–968, New York, NY, USA, 2009. Association for Computing Machinery.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine, 30(4):98–111, 2013.

P. Speckman. Kernel smoothing in partial linear models. Journal of the Royal Statistical Society. Series B (Methodological), 50(3):413–436, 1988.

B. Spinath, J. Stiensmeier-Pelster, C. Schoene, and O. Dickhäuser. Skalen zur Erfassung der Lern- und Leistungsmotivation: SELLMO. Hogrefe Verlag, Bern, 2002.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. Statistical Science, 5(4):465–472, 1990.

M.-L. Spohn, L. Henckel, and M. H. Maathuis. Identification and estimation of treatment effects on networks with interference and confounding using graphical models, 2023. Forthcoming on arXiv.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. Bernoulli, 22(3):1839–1893, 2016.

C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe. Integration in emerging social networks explains academic failure and success. Proceedings of the National Academy of Sciences, 116(3):792–797, 2019.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. Econometrica, 65(3):557–586, 1997.

C. Steglich, T. A. B. Snijders, and M. Pearson. Dynamic networks and behavior: Separating selection from influence. Sociological Methodology, 40(1):329–393, 2010.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory, volume 6, pages 583–603. University of California Press, 1972.

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

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

R. Summers. A capital intensive approach to the small sample properties of various simultaneous equation estimators. Econometrica, 33(1):1–41, 1965.

M. Taavoni and M. Arashi. High-dimensional generalized semiparametric model for longitudinal data. Statistics, 55(4):831–850, 2021a.

M. Taavoni and M. Arashi. Kernel estimation in semiparametric mixed effect longitudinal modeling. Statistical Papers, 62(3):1095–1116, 2021b.

M. Taavoni, M. Arashi, W.-L. Wang, and T.-I. Lin. Multivariate t semiparametric mixed-effects model for longitudinal data with multiple characteristics. Journal of Statistical Computation and Simulation, 91(2):260–281, 2021.

K. Takeuchi and K. Morimune. Third-order efficiency of the extended maximum likelihood estimators in a simultaneous equation system. Econometrica, 53 (1):177–200, 1985.

Y. Tang, D. Sinha, and D. Pati. Bayesian partial linear model for skewed longitudinal data. Biostatistics, 16(3):441–453, 2015.

S. J. Taylor and D. Eckles. Randomized experiments to detect and estimate social influence in networks. In S. Lehmann and Y.-Y. Ahn, editors, Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks, pages 289–322. Springer International Publishing, Cham, 2018.

E. J. Tchetgen Tchetgen, I. R. Fulcher, and I. Shpitser. Auto-g-computation of causal effects on a network. Journal of the American Statistical Association, 116(534):833–844, 2021.

H. Theil. Repeated least-squares applied to complete equation systems. Central Planning Bureau, The Hague, 1953a. Mimeographed memorandum.

H. Theil. Estimation and simultaneous correlation in complete equation systems. Central Planning Bureau, The Hague, 1953b. Mimeographed memorandum.

H. Theil. Economic forecasts and policy, volume 15 of Contributions to economic analysis. North-Holland Publishing Company, Amsterdam, 2 edition, 1961.

J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager. grf: Generalized Random Forests, 2022. URL `https://CRAN.R-project.org/package=grf`. R package version 2.1.0.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996.

P. Toulis, A. Volfovsky, and E. M. Airoldi. Estimating causal effects when treatments are entangled by network dynamics, 2021. URL `https://www.ptoulis.com/working-papers`. Working paper.

J. Ugander, B. Karrer, L. Backstrom, and J. Kleinberg. Graph cluster randomization: network exposure to multiple universes, 2013. Preprint arXiv:1305.6979.

H. Umegaki and A. Bharucha-Reid. Banach space-valued random variables and tensor products of banach spaces. Journal of Mathematical Analysis and Applications, 31(1):49–67, 1970.

M. van der Laan. Causal inference for a population of causally connected units. Journal of Causal Inference, 2(1):13–74, 2014.

M. J. van der Laan and J. M. Robins. Unified methods for censored longitudinal data and causality. Springer series in statistics. Springer, New York, 2003.

M. J. van der Laan and S. Rose. Targeted Learning. Springer Series in Statistics. Springer, New York, 2011.

M. J. van der Laan and S. Rose. Targeted Learning in Data Science. Springer Series in Statistics. Springer, New York, 2018.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1), 2006.

A. W. van der Vaart. Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

T. J. VanderWeele and E. J. Tchetgen Tchetgen. Effect partitioning under interference in two-stage randomized vaccine trials. Statistics & Probability Letters, 81(7):861–869, 2011. Statistics in Biological and Medical Sciences.

T. J. Vanderweele, G. Hong, S. M. Jones, and J. L. Brown. Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. Journal of the American Statistical Association, 108 (502):469–482, 2013.

G. Verbeke and G. Molenberghs. Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. Springer, New York, 2000.

E. F. Vonesh and V. M. Chinchilli. Linear and nonlinear models for the analysis of repeated measurements, volume 154 of Statistics: Textbooks and Monographs. Chapman & Hall/CRC, Boca Raton, Florida, 1997.

A. Vörös, Z. Boda, T. Elmer, M. Hoffman, K. Mepham, I. J. Raabe, and C. Stadtfeld. The swiss studentlife study: Investigating the emergence of an undergraduate community through dynamic, multidimensional social network data. Social Networks, 65:71–84, 2021.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests, 2017.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.

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

H. M. Wagner. A monte carlo study of estimates of simultaneous linear structural equations. Econometrica, 26(1):117–133, 1958.

N. Wang, R. J. Carroll, and X. Lin. Efficient semiparametric marginal estimation for longitudinal/clustered data. Journal of the American Statistical Association, 100(469):147–157, 2005.

Y. Wang. Causal inference under temporal and spatial interference, 2021. Preprint arXiv:2106.15074.

D. J. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393:440–442, 1998.

S. Wood and F. Scheipl. gamm4: Generalized Additive Mixed Models using "mgcv" and "lme4", 2020. URL `https://CRAN.R-project.org/package=gamm4`. R package version 0.2-6.

J. M. Wooldridge. Introductory econometrics: A modern approach. South-Western Cengage Learning, Mason, OH, 5 edition, 2013.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77 (1):1–17, 2017.

F. Yao. Efficient semiparametric instrumental variable estimation under conditional heteroskedasticity. Journal of Quantitative Economics, 10(1):32–55, 2012.

M. Yuan and D.-X. Zhou. Minimax optimal rates of estimation in high-dimensional additive models. The Annals of Statistics, 44(6):2564–2593, 2016.

S. L. Zeger and P. J. Diggle. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 50(3): 689–699, 1994.

C. Zhang, K. Mohan, and J. Pearl. Causal inference with non-IID data using linear graphical models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, Advances in Neural Information Processing Systems, 2022.

D. Zhang. Generalized linear mixed models with varying coefficients for longitudinal data. Biometrics, 60(1):8–15, 2004.

D. Zhang, X. Lin, J. Raz, and M. Sowers. Semiparametric stochastic mixed models for longitudinal data. Journal of the American Statistical Association, 93(442):710–719, 1998.

J. Zhang and L. Xue. Variable selection for generalized partially linear models with longitudinal data. Evolutionary Intelligence, 15:2473–2483, 2020.

CURRICULUM VITAE

PERSONAL INFORMATION

Name	Corinne Emmenegger
Nationality	Swiss
Date of birth	July 9, 1995
Country of birth	Switzerland

EDUCATION

02/2019 – 2023	Teaching Diploma for Grammar Schools in Mathematics, ETH Zurich
04/2019 – 03/2023	PhD in Statistics, Seminar for Statistics, ETH Zurich
02/2018 – 03/2019	MSc in Mathematics with distinction, ETH Zurich
09/2014 – 09/2018	BSc in Mathematics, ETH Zurich
08/2010 – 06/2014	Bilingual Matura (English/German), Kantonsschule Kollegium Schwyz

WORK EXPERIENCE

04/2019 – 03/2023	PhD in Statistics, Seminar for Statistics, ETH Zurich
12/2022 – 01/2023	Mathematics teacher at the Kantonsschule am Burggraben, St. Gallen
08/2022 – 09/2022	Teaching internship at the Kantonsschule am Burggraben, St. Gallen
09/2021 – 12/2021	Biostatistics internship at Roche, Basel
08/2014 – 05/2020	Collaborating author of high-school mathematics books, Orell Füssli (Titles: Algebra 7/8, Algebra 9/10, Geometrie 1, Geometrie 2, Analysis)

PUBLICATION LIST

JOURNAL PAPERS

- C. Emmenegger and P. Bühlmann. Regularizing double machine learning in partially linear endogenous models. Electronic Journal of Statistics, 15(2):6461–6543, 2021

PREPRINTS

- J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Inference for the distributional random forest, 2023. Preprint on arXiv:2302.05761
- C. Emmenegger, M.-L. Spohn, T. Elmer, and P. Bühlmann. Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning, 2022. Preprint arXiv:2206.14591
- C. Emmenegger and P. Bühlmann. Plugin machine learning for partially linear mixed-effects models with repeated measurements, 2021a. Preprint arXiv:2108.13657

OTHER ARTICLES

- C. Emmenegger. Double machine learning methods: Beyond independence. In Re-thinking High-dimensional Mathematical Statistics (Workshop Report), volume 25, pages 21–23. Oberwolfach Reports, Oberwolfach, Germany, 5 2022a. URL `http://publications.mfo.de/handle/mfo/3962`. Organizers: Florentina Bunea, Robert Nowak, Alexandre Tsybakov
- C. Emmenegger. Analyzing longitudinal data with plug-in machine learning. Bulletin of the Swiss Statistical Society, 101:8, 4 2022b. URL `https://www.stat.ch/images/bulletin/pdfs/bulletin_101.pdf`. Poster presented at the Swiss Statistical Seminar 2022, March 25

- C. Emmenegger and P. Bühlmann. Regularized double machine learning in partially linear models with unobserved confounding. In Proceedings of 63rd ISI World Statistics Congress, 2021b. URL `https://www.isi-web.org/files/docs/papers-and-abstracts/88-day2-cps020-regularized-double-machine-lea.pdf`

SOFTWARE

- D. Carl, C. Emmenegger, W. Yuan, M. Zheng, and Z. Guo. TSCI: Tools for causal inference with possibly invalid instrumental variables, 2022. URL `https://cran.r-project.org/web/packages/TSCI/index.html`. R-package available on CRAN
- C. Emmenegger. dmlalg: Double machine learning algorithms, 2021. URL `https://cran.r-project.org/web/packages/dmlalg/index.html`. R-package available on CRAN

TALKS

- C. Emmenegger. Double machine learning methods: Beyond independence. Oberwolfach Workshop "Re-thinking High-dimensional Mathematical Statistics", May 2022a
- C. Emmenegger. Double machine learning in partially linear models. Bernoulli's Round table, Department of Mathematics and Computer Science, University of Basel, December 2021
- C. Emmenegger. Regularized double machine learning in partially linear models with unobserved confounding. Bernoulli-IMS 10th World Congress in Probability and Statistics, July 2021
- C. Emmenegger. Regularized double machine learning in partially linear models with unobserved confounding. ISI 63rd World Statistics Congress, June 2021
- C. Emmenegger. Mehr Glück im Spiel dank Statistik (engl.: more luck in game thanks to statistics). Känguru Schweiz, Kangaroo goes Science

day, spring 2013

- C. Emmenegger. Regularizing double machine learning in partially linear endogenous models. Copenhagen Causality Lab (CoCaLa), Department of Mathematical Sciences, University of Copenhagen, April 2021

POSTERS

- C. Emmenegger. Machine learning for repeated measurements data. Women in Data Science, May 2022b
- C. Emmenegger. Analyzing longitudinal data with plug-in machine learning. Swiss Statistics Seminar, March 2022