# Causal Machine Learning for Moderation Effects

Nora Bearth* & Michael Lechner†

**Abstract**

It is valuable for any decision maker to know the impact of decisions (treatments) on average and for subgroups. The causal machine learning literature has recently provided tools for estimating group average treatment effects (GATE) to understand treatment heterogeneity better. This paper addresses the challenge of interpreting such differences in treatment effects between groups while accounting for variations in other covariates. We propose a new parameter, the *balanced group average treatment effect* (BGATE), which measures a GATE with a specific distribution of a priori-determined covariates. By taking the difference of two BGATEs, we can analyze heterogeneity more meaningfully than by comparing two GATEs. The estimation strategy for this parameter is based on double/debiased machine learning for discrete treatments in an unconfoundedness setting, and the estimator is shown to be $\sqrt{N}$-consistent and asymptotically normal under standard conditions. Adding additional identifying assumptions allows specific balanced differences in treatment effects between groups to be interpreted causally, leading to the *causal balanced group average treatment effect*. We explore the finite sample properties in a small-scale simulation study and demonstrate the usefulness of these parameters in an empirical example.

**JEL classification:** C14, C21

**Keywords:** Causal machine learning, double/debiased machine learning, treatment effect heterogeneity, moderation effects, causal moderation

*Swiss Institute for Empirical Economic Research of the University of St. Gallen (SEW-HSG), Varnbüelstrasse 14, 9000 St. Gallen, CH, E-mail: nora.bearth@unisg.ch

†Swiss Institute for Empirical Economic Research of the University of St. Gallen (SEW-HSG), Varnbüelstrasse 14, 9000 St. Gallen, CH, E-mail: michael.lechner@unisg.ch, Michael Lechner is also affiliated with CEPR, London, CESIfo, Munich, IAB, Nuremberg and IZA, Bonn.

Financial support from the Swiss National Science Foundation (SNSF) is gratefully acknowledged. The study is part of the project "Chances and risks of data-driven decision making for labour market policy" (grant number SNSF 407740_187301) of the Swiss National Research Program "Digital Transformation" (NRP 77). We thank Daniele Ballinari, Hannah Busshoff, Jonathan Chassot, Riccardo Di Francesco, and Jana Mareckova for comments and suggestions on a previous version of this paper. The paper was presented at the annual meeting of the Verein für Socialpolitik (2023) and at the University of St. Gallen. We thank participants for helpful comments and suggestions. Last, we thank GPT-3.5 and grammarly for editorial support.

# 1 Introduction

Detecting and interpreting heterogeneity in treatment effects is crucial for understanding the impact of interventions and (policy) decisions. Researchers have recently developed many methods to estimate heterogeneous treatment effects. However, there are still limitations in interpreting these effects. For example, suppose that the effect of a particular training program for the unemployed is larger for women than for men. By comparing the average treatment effects for these two groups, without taking into account the different distribution of other covariates of men and women, such as education or labor market experience, we might implicitly compare a group with longer labor market experience (men) with a group with shorter labor market experience (women). Therefore, obtaining a balanced distribution of relevant characteristics across different groups may be crucial to ensure proper comparisons and draw meaningful conclusions. In the specific example, we might want to ensure that both groups have the same average years of labor market experience. This approach allows us to isolate the differences in causal effects between groups in a way that is not confounded by certain other covariates. However, the difference may still be due to differences in unobservable characteristics that vary between the two groups of interest. In addition, determining whether a particular variable causes differences in treatment effects is often of interest. In the literature, these variables are called causal moderator variables. However, to interpret the differences in treatment effects causally, i.e. in such a way that the group variable can be considered an unconfounded moderator, it is crucial not only to have balanced distributions of all covariates that confound the moderation effect across groups but also to check that additional assumptions hold.

This paper discusses how to estimate differences in treatment effects between groups in an unconfoundedness setting. First, a new parameter called *balanced group average treatment effect (BGATE)* is introduced. This parameter is a group average treatment effect (GATE) with a specific distribution of a-priori-determined covariates. It is beneficial for comparing the treatment effects of two groups with each other. This results in the difference of two BGATEs called $\Delta$BGATE. We demonstrate how it relates to a difference of two group average treatment effects ($\Delta$GATE), discuss its identification and propose an estimator for discrete moderators[1] and discrete treatment effects based on double/debiased machine learning (subsequently abbreviated as DML) (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, & Robins, 2018).[2] Fur-

[1]For simplicity, we refer to the variable for which we find heterogeneous treatment effects as moderator variable.

[2]For better readability we use a binary treatment and a binary moderator in the main body of the paper. All the proofs in the Appendix are done for a discrete treatment and moderator variable.

thermore, we show that the estimator is asymptotically normal centred at the true value, which allows valid inference. A small-scale simulation study provides evidence that the estimator performs well in finite samples and an empirical example using administrative labor market data from Switzerland demonstrates its practicality. Finally, we show which conditions are needed to interpret a ΔBGATE causally. For this purpose, a second parameter, namely the *causal balanced group average treatment effect* (ΔCBGATE)[3] is introduced. The ΔCBGATE can be interpreted causally under additional assumptions. The proposed DML estimator is asymptotically normal centred at the true value, which allows valid inference and its variance achieves the semi-parametric efficiency bound (Hahn, 1998).

The paper proceeds as follows: In Section 2, we define the parameter of interest for the case of a binary treatment and a binary moderator and discuss its identification under unconfoundedness. Section 3 proposes an estimation strategy based on DML and shows its asymptotic properties. Section 4 depicts the design and results of a small-scale Monte Carlo simulation. In Section 5, we demonstrate how the new parameter can be used in practice. Finally, in Section 6, we discuss the causal moderation estimator and explain the additional identifying assumptions that must be met. Last, Section 7 concludes.

### 1.1 Related Literature

Several authors have recently contributed to the topic of effect heterogeneity (e.g., Tian, Alizadeh, Gentles, & Tibshirani, 2014; Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019; Künzel, Sekhon, Bickel, & Yu, 2019; Nie & Wager, 2021; Semenova & Chernozhukov, 2021; Knaus, 2022; Di Francesco, 2023; Foster & Syrgkanis, 2023; Kennedy, 2023). The proposed methods make it possible to estimate heterogeneities between fine-grained subgroups more accurately. For a recent review of the various methods and their performance, see Knaus, Lechner, & Strittmatter (2021). As a result, many applied papers now use such methods to detect heterogeneities (e.g., Davis & Heller, 2017; Knaus, Lechner, & Strittmatter, 2022; Cockx, Lechner, & Bollens, 2023). However, decision makers are often more interested in heterogeneities for a small subset of covariates than at the finest possible granularity. Therefore, several papers show how to detect (low-dimensional) heterogeneities at the group level, called "Group Average Treatment Effects" (GATEs).[4] Several approaches have been developed to estimate GATEs. Abrevaya, Hsu, & Lieli (2015) show how

[3]To keep the notation consistent, we call the parameter ΔCBGATE, although we can only claim causality by taking the difference.

[4]A GATE is a conditional average treatment effect (CATE) with a small number of conditional variables, often only a single one.

to identify GATEs nonparametrically under unconfoundedness and estimate them using inverse probability weighting estimators. Athey, Tibshirani, & Wager (2019) and Lechner & Mareckova (2022) modify a random forest algorithm to adjust for confounding to estimate heterogeneous treatment effects. Semenova & Chernozhukov (2021) use the DML framework to find heterogeneity based on linear models. Zimmert & Lechner (2019) and Fan, Hsu, Lieli, & Zhang (2022) develop a two-step estimator that allows estimating GATEs nonparametrically. The first step is estimated using machine learning methods, and in the second step, they apply a nonparametric local constant regression. We add to this literature by proposing a new parameter of interest: the difference between two GATEs with (partially) balanced characteristics. By taking the difference, we can disentangle the difference in treatment effects from the difference in the distribution of covariates between the two groups. Chernozhukov, Fernández-Val, & Luo (2018) look at the problem of interpreting heterogeneous treatment effects from another angle. They suggest sorting the estimated partial effects by percentiles and comparing the covariate means of observations falling in the different percentiles to see which individuals with which characteristics are most positively and negatively affected. In contrast, we are not interested in finding the characteristics of the individuals with the most positive and negative effects but in comparing the treatment effects of the two groups while balancing the distribution of some covariates.

As has already become apparent from the discussion of the first part of the relevant literature, the DML literature is closely related to our work. Chernozhukov et al. (2018) developed this framework, which allows using machine learning methods for causal analysis. Machine learning algorithms may introduce two biases: a regularisation and an overfitting bias. If the results are biased, it is impossible to make causal inference. The main idea of DML is that by using Neyman-orthogonal score functions, we can overcome the regularisation bias, and by using cross-fitting, an efficient form of sample splitting, we can overcome the overfitting bias. Based on this paper, many papers have adapted the framework for different settings, for example, for continuous treatments (e.g., Kennedy, Ma, McHugh, & Small, 2017; Semenova & Chernozhukov, 2021), for mediation analysis (Farbmacher, Huber, Laffērs, Langen, & Spindler, 2022), for panel data (e.g. Clarke & Polselli, 2023) or for difference-in-difference estimation (e.g., Zimmert, 2018; Sant'Anna & Zhao, 2020). We add to this literature by using this highly flexible framework to estimate the new parameters of interest.

Last, we add to the literature on causal moderation by showing what additional assumptions must be met to interpret the difference in treatment effects causally. Discussions of moderation effects can be found primarily in the psychology and political science literature (e.g., Gogineni, Alsup, & Gillespie, 1995; Frazier, Tix, & Barron, 2004; Rose, Holmbeck, Coakley, & Franks, 2004; Dearing & Hamilton, 2006; Fairchild & McQuillin, 2010; Marsh, Hau, Wen, Nagengast, & Morin, 2013; Bansak, Bechtel, & Margalit, 2021; Blackwell & Olson, 2022). Common approaches to analyzing moderation effects are interaction effects in a regression or subgroup analysis. Without additional assumptions, these effects cannot be interpreted causally. Bansak (2021) studies in an experimental setting causal moderation effects by showing what identifying assumptions are needed. Because of the randomisation of treatment, he shows that it is possible to estimate the causal effect of the moderator on the outcome separately for the treated and control units and then subtract both estimates to obtain the moderator effect. However, such differences in effects cannot be interpreted causally if the moderator variable influences some covariates. Moderators, such as gender or education, often influence other covariates. In this scenario, a causal interpretation is also impossible in an experiment. In addition, Bansak & Nowacki (2022) have recently shown how to identify and estimate subgroup differences in a regression discontinuity design. They introduce two different parameters. One is the difference between two treatment effects when a moderator variable causes the difference, and the other when the difference is only associated with that variable. They propose a local linear regression, with and without a matched sample, to estimate their new causal moderator effect.

# 2 Effects of Interest

## 2.1 Definition

The causal moderation framework used in this paper is based on the potential outcome framework of Rubin (1974). A causal effect is defined as the difference between two potential outcomes, whereas for a unit, we only observe one of these potential outcomes. Therefore, finding a credible counterfactual is problematic.

We observe $N$ i.i.d. observations of the independent random variables $H_i = (D_i, Y_i, Z_i, X_i)$ according to an unknown probability distribution $\mathbb{P}$. Here, the focus is on a treatment $D_i$ and a moderator $Z_i$ which, for simplicity, are assumed to be binary.[5] As usual, the potential outcomes are indexed by the treatment variable: $(Y_i^0, Y_i^1)$. Finally, a set of $k \in \{1, \ldots, p\}$ covariates $X_{i,k}$ might simultaneously affect treatment allocation and potential outcomes, where

[5] The realizations of the treatment variable are $d \in \{0, 1\}$, and of the moderator variable $z \in \{0, 1\}$. The formal theory presented in Appendix A is based on fixed numbers of discrete values for $Z_i$ and $D_i$.

$X_i = (X_{i,1}, \ldots, X_{i,p})$.

Since we only observe realizations of one of the potential outcomes, we can never consistently estimate realizations of the individual treatment effect (ITE) $\xi_i = Y_i^1 - Y_i^0$. However, under suitable assumptions, the identification of, for example, the average treatment effect (ATE) $\theta = \mathrm{E}[Y_i^1 - Y_i^0]$ is possible (Imbens & Wooldridge, 2009). It is often interesting to additionally investigate different aspects of the heterogeneity of the $\xi_i$ which can be captured by so-called conditional average treatment effects (CATE). A CATE measures the average treatment effect conditional on a (sub-) set of covariates $X_i$. The individualized average treatment effect (IATE) and the group average treatment effect (GATE) are specific CATEs. The IATE measures the treatment effect at the most granular aggregation level. Namely, it compares the average effect of the treatment for all individuals with a specific value of all relevant covariates used. Formally, the IATE is defined as follows:

$$\tau(x, z) = \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = z, X_i = x]$$

The GATE measures the treatment effect at the group level, i.e. at a more aggregated level than the IATE, but still at a finer level than the ATE. Formally, the GATE is defined as follows:

$$\theta^G(z) = \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = z] = \mathrm{E}[\tau(X_i, z) | Z_i = z]$$

As long as the interest lies only in describing effect heterogeneity, IATEs and GATEs are sufficient. However, if the interest lies in the difference in treatment effects between the two groups, the difference between the two GATEs ($\Delta$GATE)

$$\begin{aligned} \theta^{\Delta G} &= \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 1] - \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 0] \\ &= \mathrm{E}[\tau(X_i, 1) | Z_i = 1] - \mathrm{E}[\tau(X_i, 0) | Z_i = 0] \end{aligned}$$

may be difficult to interpret because the two groups may differ in the distribution of other covariates $X_i$.

Thus, we introduce a new parameter, the *balanced group average treatment effect* (BGATE). The variables used to balance the GATEs are denoted as $W_i$. $W_i$ is part of $X_i$. If $W_i$ is empty, or $W_i$ is independent of $Z_i$, the BGATE reduces to the GATE. Thus, the new parameter of interest, denoted by $\theta^B(z)$, is defined as

$$\theta^B(z) = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0 | Z_i = z, W_i]] = \mathrm{E}[\mathrm{E}[\tau(X_i, z) | Z_i = z, W_i]]$$

and its difference as $\theta^{\Delta B}$ as

$$\begin{aligned}\theta^{\Delta B} &= \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 1, W_i] - \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 0, W_i]] \\ &= \mathrm{E}[\mathrm{E}[\tau(X_i, 1) | Z_i = 1, W_i] - \mathrm{E}[\tau(X_i, 0) | Z_i = 0, W_i]]\end{aligned}$$

A $\Delta$BGATE ($\theta^{\Delta B}$) represents the difference between two groups, adjusting the distribution of some other covariates ($W_i$) in both groups to the overall population distribution. The $\Delta$BGATE usually shows associational moderation effects.

## 2.2 Identification

To identify the GATE, BGATE, $\Delta$GATE or $\Delta$BGATE in an unconfoundedness setting, usual identifying assumptions are needed (e.g., Imbens, 2004):

**Assumption 1.** (Conditional Independence (CIA))

$$(Y_i^1, Y_i^0) \perp D_i | X_i = x, Z_i = z, \quad \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}$$

The CIA assumption requires the potential outcomes $(Y_i^1, Y_i^0)$ to be independent of the treatment assignment ($D_i$) for given values of the confounding variables $(X_i, Z_i)$.

**Assumption 2.** (Common support (CS))

$$0 < P(D_i = d | X_i = x, Z_i = z) < 1, \quad \forall d \in \{0, 1\}, \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}$$

For any given values of $X_i$ and $Z_i$, there must be an observation for each treatment state $d \in \{0, 1\}$.

**Assumption 3.** (Exogenity of confounders)

$$X_i^0 = X_i^1, \quad Z_i^0 = Z_i^1$$

*where $X_i^d$ and $Z_i^d$ are potential variables that depend on the treatment.*

**Assumption 4.** (Stable Unit Treatment Value Assumption (SUTVA))

$$Y_i = D_i Y^1 + (1 - D_i) Y^0$$

SUTVA requires that there are no unrepresented treatments in the population of interest (everyone is assigned to a treatment state) and that treatment assignment does not change the effects.

**Lemma 1.** *Under Assumptions 1 to 4, the parameter* $\theta^B(z) = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0 | Z_i = z, W_i]]$ *is identified as* $\mathrm{E}[\mathrm{E}[\mu_1(Z_i, X_i) - \mu_0(Z_i, X_i) | Z_i = z, W_i]]$ *with* $\mu_d(z, x) = \mathrm{E}[Y_i | D_i = d, Z_i = z, X_i = x]$*. Hence,* $\theta^{\Delta B} = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 1, W_i] - \mathrm{E}[Y_i^1 - Y_i^0 | Z_i = 0, W_i]]$ *is identified as* $\mathrm{E}[\mathrm{E}[\mu_1(Z_i, X_i) - \mu_0(Z_i, X_i) | Z_i = 1, W_i] - \mathrm{E}[\mu_1(Z_i, X_i) - \mu_0(Z_i, X_i) | Z_i = 0, W_i]]$.

For the proof of Lemma 1 see Appendix A.2.1.

# 3 Estimation and Inference

## 3.1 Estimator

Since we are interested in differences in treatment effects, the estimation strategy focuses on the ΔBGATE. However, the estimator can easily be adapted to estimate the BGATE. The identification results suggest a three-step estimation strategy. To obtain a flexible estimator that allows for a potentially high-dimensional vector of covariates, we rely on the methodology of DML as proposed by Chernozhukov et al. (2018). In the first step, the usual double robust score function is estimated. In the second step, the score function is regressed on the two indicator variables defined by the different values of the moderator variable $Z_i$ and the covariates we want to balance $W_i$. Last, we take the difference between the two groups defined by the moderator and average over the variables we balance. This approach is close to the approach of Kennedy (2023) for estimating CATEs with the DR-learner. However, instead of estimating a CATE, we average over $W_i$.

The estimated doubly robust score function is given by the following expression:

$$\hat{\phi}^{\Delta B}(h; \theta^{\Delta B}, \hat{\eta}) = \hat{g}_1(w) - \hat{g}_0(w) + \frac{z\left(\hat{\delta}(h) - \hat{g}_1(w)\right)}{\hat{\lambda}_1(w)} - \frac{(1-z)\left(\hat{\delta}(h) - \hat{g}_0(w)\right)}{1 - \hat{\lambda}_1(w)} - \theta^{\Delta B}$$

with

$$\begin{aligned}
\delta(h) &= \mu_1(z,x) - \mu_0(z,x) + \frac{d(y-\mu_1(z,x))}{\pi_1(z,x)} - \frac{(1-d)(y-\mu_0(z,x))}{1-\pi_1(z,x)} \\
\lambda_z(w) &= P(Z_i = z|W_i = w) \\
\mu_d(z,x) &= \mathrm{E}[Y_i|D_i = d, Z_i = z, X_i = x] \\
\pi_d(z,x) &= P(D_i = d|Z_i = z, X_i = x) \\
g_z(w) &= \mathrm{E}[\delta(H_i)|Z_i = z, W_i = w].
\end{aligned}$$

$\hat{\delta}(h)$, $\hat{\mu}_d(z,x)$, $\hat{\lambda}_z(w)$ and $\hat{\pi}_d(z,x)$ denote the estimated values of $\delta(h)$, $\mu_d(z,x)$, $\lambda_z(w)$ and $\pi_d(z,x)$, respectively. Furthermore, notice that $\hat{g}_z(w) = \hat{\mathrm{E}}[\hat{\delta}(H_i)|Z_i = z, W_i = w]$ is the regression of $\hat{\delta}(H_i)$ on $Z_i$ and $W_i$ and that $\tilde{g}_z(w) = \hat{\mathrm{E}}[\delta(H_i)|Z_i = z, W_i = w]$ is the corresponding oracle regression of $\delta(H_i)$ on $Z_i$ and $W_i$. Hence, $\hat{\mathrm{E}}[\ldots|\ldots]$ denotes a generic regression estimator. Last, the estimated nuisance parameters are $\hat{\eta} = (\hat{\mu}_d(z,x), \hat{\pi}_d(z,x), \hat{\lambda}_z(w), \hat{g}_z(w))$.

As explained above, the score function $\hat{\delta}(h)$ has to be estimated. The product of the nuisance function errors for $\hat{\delta}(h)$ must converge faster than or equal to $\sqrt{N}$, and cross-fitting with K-folds ($K > 1$) must be used. In the second estimation step, the product of the nuisance function errors has to converge with $\sqrt{N}$ and cross-fitting with J-folds ($J > 1$) has to be used.[6] Then, the estimator is $\sqrt{N}$-consistent and asymptotically normal (see Subsection 3.3). Because $E[\phi^{\Delta B}(H_i;\theta^{\Delta B},\eta)] = 0$, the variance of $\hat{\theta}^{\Delta B}$ is given by

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}^{\Delta B}) &= \mathrm{Var}(\phi^{\Delta B}(H_i;\theta^{\Delta B},\eta)) \\
&= \mathrm{E}[\phi^{\Delta B}(H_i;\theta^{\Delta B},\eta)^2] - \underbrace{\mathrm{E}[\phi^{\Delta B}(H_i;\theta^{\Delta B},\eta)]^2}_{=0} \\
&= \mathrm{E}[\phi^{\Delta B}(H_i;\theta^{\Delta B},\eta)^2]
\end{aligned}$$

and is estimated by

$$\widehat{\mathrm{Var}}(\hat{\theta}^{\Delta B}) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in S_j}[\hat{\phi}^{\Delta B}(H_i;\theta^{\Delta B},\hat{\eta})]^2.$$

[6] For example, random forests (Wager & Walther, 2015; Syrgkanis & Zampetakis, 2020), Lasso (Belloni & Chernozhukov, 2013), boosting (Luo, Spindler, & Kück, 2016) or neural nets (Farrell, Liang, & Misra, 2021) converge at a rate faster than or equal to $N^{1/4}$ (under certain conditions). Therefore, these machine learners can be used to estimate the nuisance parameters.

### 3.2 Implementation

For the implementation of the estimator, the weights (e.g., $\frac{z}{\lambda_1(w)}$) are normalized to ensure that they do not explode. Furthermore, the weights are truncated such that the weight of each observation is no more than five per cent of the sum of all weights. In a third step, we renormalize the (possibly) truncated weights so that they sum to one (Huber, Lechner, & Wunsch, 2013). For propensity scores too close to 0 or 1, unnormalized and untruncated weights can lead to implausibly large effect estimates (Busso, DiNardo, & McCrary, 2014). See Algorithm 2 in Appendix B for the exact procedure of normalizing and truncating the propensity scores.

Algorithm 1 is written without normalizing and truncating the propensity score. However, all propensity scores, namely $\lambda_z(w)$ and $\pi_d(z, x)$ are normalized and truncated.

**Algorithm 1:** DML FOR ΔBGATE

**Input** : Data: $h_i = \{x_i, z_i, d_i, y_i\}$

**Output:** $\hat{\theta}^{\Delta B}$, SE($\hat{\theta}^{\Delta B}$)

**begin**

**create folds:** Split sample into K random folds $(S_k)_{k=1}^K$ of observations $\{1, \ldots, \frac{N}{K}\}$ with size of each fold $\frac{N}{K}$. Define $S_k^c := \{1, \ldots, N\} \backslash \{S_k\}$

**for** $k$ in $\{1, \ldots, K\}$ **do**

RESPONSE FUNCTIONS:

estimate:

$\hat{\mu}_1(z_i, x_i) = \hat{\text{E}}[Y_i|D_i = 1, X_i = x_i, Z_i = z_i]$ in $\{x_i, y_i, z_i\}_{i \in S_k^c, d_i=1}$

$\hat{\mu}_0(z_i, x_i) = \hat{\text{E}}[Y_i|D_i = 0, X_i = x_i, Z_i = z_i]$ in $\{x_i, y_i, z_i\}_{i \in S_k^c, d_i=0}$

PROPENSITY SCORE:

estimate: $\hat{\pi}_1(x_i, z_i) = \hat{P}(D_i = 1|X_i = x_i, Z_i = z_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

PSEUDO-OUTCOME:

estimate: $\hat{\delta}(h_i) = \hat{\mu}_1(z_i, x_i) - \hat{\mu}_0(z_i, x_i) + \frac{d_i(y_i - \hat{\mu}_1(z_i, x_i))}{\hat{\pi}_1(x_i, z_i)} - \frac{(1-d_i)(y_i - \hat{\mu}_0(z_i, x_i))}{1-\hat{\pi}_1(x_i, z_i)}$ in $\{h_i\}_{i \in S_k}$

**end**

**create folds:** Split sample $S_k$ into J random folds $(S_j)_{j=1}^J$ of observations $\{1, \ldots, \frac{N}{K \cdot J}\}$ with size of each fold $\frac{N}{K \cdot J}$. Define $S_j^c := \{1, \ldots, \frac{N}{K}\} \backslash \{S_j\}$

**for** $j$ in $\{1, \ldots, J\}$ **do**

PSEUDO-OUTCOME REGRESSION:

estimate:

$\hat{g}_1(w_i) = \hat{\text{E}}[\hat{\delta}(h_i)|Z_i = 1, W_i = w_i]$ in $\{h_i\}_{i \in S_j^c, z_i=1}$

$\hat{g}_0(w_i) = \hat{\text{E}}[\hat{\delta}(h_i)|Z_i = 0, W_i = w_i]$ in $\{h_i\}_{i \in S_j^c, z_i=0}$

PROPENSITY SCORE:

estimate: $\hat{\lambda}_1(w_i) = \hat{P}(Z_i = 1|W_i = w_i)$ in $\{w_i, z_i\}_{i \in S_j^c}$

ΔBGATE FUNCTION:

EFFECT:

estimate: $\hat{\theta}^{\Delta B} = \frac{K \cdot J}{N} \sum_{i \in S_j} \Big[\hat{g}_1(w_i) - \hat{g}_0(w_i) + \frac{z_i(\hat{\delta}(h_i) - \hat{g}_1(w_i))}{\hat{\lambda}_1(w_i)} - \frac{(1-z_i)(\hat{\delta}(h_i) - \hat{g}_0(w_i))}{1-\hat{\lambda}_1(w_i)}\Big]$

STANDARD ERRORS:

estimate:

$\hat{\theta}^{\Delta B_{SE}} = \frac{K \cdot J}{N} \sum_{i \in S_j} \Big[\Big(\hat{g}_1(w_i) - \hat{g}_0(w_i) + \frac{z_i(\hat{\delta}(h_i) - \hat{g}_1(w_i))}{\hat{\lambda}_1(w_i)} - \frac{(1-z_i)(\hat{\delta}(h_i) - \hat{g}_0(w_i))}{1-\hat{\lambda}_1(w_i)}\Big)^2\Big]$

**end**

estimate effect: $\hat{\theta}^{\Delta B} = \frac{1}{J \cdot K} \sum_{j=1}^J \sum_{k=1}^K \hat{\theta}_{j,k}^{\Delta B}$

estimate standard errors: $SE(\hat{\theta}^{\Delta B}) = \sqrt{\frac{1}{J \cdot K} \sum_{j=1}^J \sum_{k=1}^K \hat{\theta}_{j,k}^{\Delta B_{SE}} - \left(\hat{\theta}_{j,k}^{\Delta B}\right)^2}$

**end**

### 3.3 Asymptotic Properties

We investigate the asymptotic properties of the estimator. The following assumptions are imposed:

**Assumption 5.** (Overlap)

*The propensity scores* $\lambda_z(w)$ *and* $\pi_d(z,x)$ *are bounded away from 0 and 1:*

$$\kappa < \lambda_z(w), \pi_d(z,x) < 1-\kappa \quad \forall x \in \mathcal{X}, z \in \mathcal{Z},$$

*for some* $\kappa > 0$.

**Assumption 6.** (Consistency)

*The estimators of the nuisance functions are sup-norm consistent:*

$$\sup_{x\in\mathcal{X}, z\in\mathcal{Z}} |\hat{\mu}_d(z,x) - \mu_d(z,x)| \xrightarrow{p} 0$$
$$\sup_{x\in\mathcal{X}, z\in\mathcal{Z}} |\hat{\pi}_d(z,x) - \pi_d(z,x)| \xrightarrow{p} 0$$
$$\sup_{w\in\mathcal{W}} |\hat{\lambda}_z(w) - \lambda_z(w)| \xrightarrow{p} 0$$
$$\sup_{w\in\mathcal{W}} |\tilde{g}_z(w) - g_z(w)| \xrightarrow{p} 0$$

**Assumption 7.** (Risk decay)

*The products of the estimation errors for the outcome and propensity models decay as*

$$\mathrm{E}\left[(\hat{\mu}_d(Z_i,X_i) - \mu_d(Z_i,X_i))^2\right] \mathrm{E}\left[(\hat{\pi}_d(Z_i,X_i) - \pi_d(Z_i,X_i))^2\right] = o_p\left(\frac{1}{N}\right)$$
$$\mathrm{E}\left[(\tilde{g}_z(W_i) - g_z(W_i))^2\right] \mathrm{E}\left[(\hat{\lambda}_z(W_i) - \lambda_z(W_i))^2\right] = o_p\left(\frac{1}{N}\right)$$

*If both nuisance parameters are estimated with the parametric ($\sqrt{N}$-consistent) rate, then the product of the errors would be bounded by* $O_p\left(\frac{1}{N^2}\right)$. *Hence, it is sufficient for the estimators of the nuisance parameters to be* $N^{1/4}$*-consistent.*

**Assumption 8.** (Boundness of conditional variances)

*The conditional variances of the outcomes and score functions are bounded:*

$$\sup_{w \in \mathcal{W}} \mathrm{Var}(\delta(H_i)|Z_i = z, W_i = w) < \epsilon_{z1} < \infty$$

$$\sup_{w \in \mathcal{W}} \mathrm{Var}(\hat{\delta}(H_i)|Z_i = z, W_i = w) < \epsilon_{z0} < \infty$$

$$\sup_{x \in \mathcal{X}, z \in \mathcal{Z}} \mathrm{Var}(Y_i|D_i = d, Z_i = z, X_i = x) < \epsilon_d < \infty$$

*for some constants* $\epsilon_{z1}, \epsilon_{z0}, \epsilon_d > 0$.

**Assumption 9.** (Stability)

*The second step regression estimator* $\hat{\mathrm{E}}[\ldots|\ldots]$ *has to be stable in the sense of Definition 1 in Appendix A.2.3 (Definition 1 in Kennedy (2023)).*

Assumptions 5 to 8 made are standard in the DML literature (Chernozhukov et al., 2018). The only difference is that these assumptions are applied for the first and the second estimation step. Assumption 9 is needed because we regress in the second estimation step an already estimated quantity $\hat{\delta}$ on some variables $Z_i$ and $W_i$. Kennedy (2023) proves that linear smoothers, such as linear regression, random forests or nearest neighbour matching are stable.

Given these assumptions, we can derive the main theoretical result:

**Theorem 1.** *Under Assumptions 5 to 9, the proposed estimation strategy for the* $\Delta BGATE$ *obeys*

$$\sqrt{N}(\hat{\theta}^{\Delta B} - \theta^{\Delta B}) \xrightarrow{d} N(0, V^\star)$$

*with* $V^\star = \mathrm{E}[\phi^{\Delta B}(H_i; \theta^{\Delta B}, \eta)^2]$

Theorem 1 states that the estimator is $\sqrt{N}$-consistent and asymptotically normal. See Appendix A.2.3 for the proof.

# 4 Simulation Study

## 4.1 Data Generating Process (DGP)

We start with simulating a p-dimensional covariate matrix $X_{i,p}$ with p=6. The first two covariates are drawn from a uniform distribution $X_{i,0}, X_{i,1} \sim \mathcal{U}[0,1]$ and the remaining covariates from a normal distribution $X_{i,2} \ldots, X_{i,p-1} \sim \mathcal{N}\left(0.5, \sqrt{1/12}\right)$. All covariates have a mean of 0.5 and a standard deviation of $\sqrt{1/12}$. The moderator variable $Z_i$ is drawn from a Bernoulli distribution with probability $P(Z_i = 1|X_{i,0}, X_{i,1}) = (0.1 + 0.8\beta(X_{i,0} \times X_{i,1}; 2, 4))$.[7] The propensity score is created similarly as in Künzel, Sekhon, Bickel, & Yu (2019) and Wager & Athey (2018). The treatment variable $D_i$ is drawn from a Bernoulli distribution with probability $P(D_i = 1|X_{i,0}, X_{i,1}, X_{i,2}, X_{i,5}, Z_i) = \left(0.2 + 0.6\beta(\frac{X_{i,0}+X_{i,1}+X_{i,2}+X_{i,5}+Z_i}{5}; 2, 4)\right)$.

This simulation study includes three variations of the DGP: a non-linear DGP with linear heterogeneity, a non-linear DGP with non-linear heterogeneity, and a non-linear DGP with non-linear heterogeneity and a confounder that is influenced by the moderator variable $Z_i$. The empirical reason for the third DGP is that researchers are often interested in moderators such as gender or education. These socio-demographic characteristics often influence other covariates used in the analysis. Therefore, the procedure must also work well in such a case. For this third DGP, $X_{i,5}$ is drawn from a Bernoulli distribution with probability $P(X_{i,5} = 1|Z_i) = \frac{1}{1+\exp(-2Z_i)}$.

Next, the response functions under treatment and non-treatment and the two states of the moderator variable are specified. The non-treatment response function is specified similarly as in Nie & Wager (2021) and creates a difficult non-linear setting. They are given by

$$\mu_0(1, X_i) = \sin(\pi \times X_{i,0} \times X_{i,1}) + (X_{i,2} - 0.5)^2 + 0.1X_{i,3} + 0.3X_{i,5}$$
$$\mu_0(0, X_i) = \sin(\pi \times X_{i,0} \times X_{i,1}) + (X_{i,2} - 0.5)^2 + 0.1X_{i,3} + 0.3X_{i,5}$$

The response functions under treatment for *linear heterogeneity* are defined as

$$\mu_1(1, X_i) = \mu_0(1, X_i) + [X_{i,0}, X_{i,1}, X_{i,2}, X_{i,5}] \times \beta_1 + 0.2Z_i$$
$$\mu_1(0, X_i) = \mu_0(0, X_i) + [X_{i,0}, X_{i,1}, X_{i,2}, X_{i,5}] \times \beta_0 + 0.5Z_i.$$

with $\beta_1 = [0.7, 0.1, 0.7, 0.4]^T$ and $\beta_0 = [0.2, 0.3, 0.6, 0.3]^T$. The response functions under treat-

[7] $\beta(X_{i,0} \times X_{i,1}; 2, 4)$ detones the cdf of a beta distribution with the shape parameters $\alpha = 2$ and $\beta = 4$.

ment for *non-linear heterogeneity* are defined differently, namely as:

$$\mu_1(1, X_i) = \mu_0(1, X_i) + \sin(4.9X_{i,0}) + \sin(2X_{i,1}) + 0.7X_{i,2}^4 + 0.4X_{i,5} + 0.2Z_i$$
$$\mu_1(0, X_i) = \mu_0(0, X_i) + \sin(1.4X_{i,0}) + \sin(6X_{i,1}) + 0.6X_{i,2}^2 + 0.3X_{i,5} + 0.5Z_i.$$

They are chosen in such a way that the ΔBGATE is different from the ΔGATE. Last, we simulate the potential outcomes as $Y_i^d(z) = \mu_d(z, X_i) + e_{i,d,z}\ \forall z \in \{0, 1\}$ with noise $e_{i,d,z} \sim \mathcal{N}(0, 1)$. Summing up, the data consists of an observable quadruple $(y_{i,r}, d_{i,r}, z_{i,r}, x_{i,r})$ and the true values are estimated on a sample with $N = 100,000$.

### 4.2 Effects of Interest and Estimators

Two ΔBGATEs and a ΔGATE are investigated, namely:

$$\theta^{\Delta G} = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0|Z_i = 1] - \mathrm{E}[Y_i^1 - Y_i^0|Z_i = 0]]$$
$$\theta_{X_0}^{\Delta B} = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0|Z_i = 1, X_{i,0}] - \mathrm{E}[Y_i^1 - Y_i^0|Z_i = 0, X_{i,0}]]$$
$$\theta_{X_2}^{\Delta B} = \mathrm{E}[\mathrm{E}[Y_i^1 - Y_i^0|Z_i = 1, X_{i,2}] - \mathrm{E}[Y_i^1 - Y_i^0|Z_i = 0, X_{i,2}]]$$

As shown in the previous subsection, $X_{i,0}$ is unbalanced across the two groups of the moderator, whereas $X_{i,2}$ is balanced. In Appendix C.2.2, several results for other effects of interest are depicted.

We estimate these effects with two different estimation approaches. One estimation strategy uses Algorithm 1 and 2 with $K = 2$ folds in the first estimation step and $J = 5$ in the second. In that case, all nuisance functions are estimated using random forests (number of trees: 1000). In the second estimation strategy, we estimate the nuisance functions in the first estimation step in the same way. As shown by Bach, Schacht, Chernozhukov, Klaassen, & Spindler (2024) it is crucial to tune the learners used for the estimation of the nuisance parameters. Table C.1 in Appendix C.2.1 show the hyperparameters used for the different random forests. Second, instead of using DML, the score function is regressed on the covariates we want to balance and the value of interest of the moderator variable. This is done by using a linear model (OLS).[8] As long as the heterogeneities are linear, this approach should work well. However, if the heterogeneities

[8]The results in Appendix C.2.2 show two different versions of using DML in the second step. Once, using a linear regression for estimating the outcome regressions, and once using a random forest. The double robust property of the DML estimator leads to less biased results for the DML with OLS version than by simply taking the difference of the two outcome regressions as shown in Figure 5 and 7 in the next section.

are non-linear, it leads to biased results.

### 4.3 Simulation Design

As shown in Section 4.1, there are three different DGPs. For each DGP, we investigate a simulation with 2,500 observations (N) with 1000 replications (R), and N $= 10,000$ with R $= 250$.[9] Furthermore, for the second estimation step, we estimate each effect of interest, either with a linear regression or with DML using random forests. DML is always used to estimate the pseudo outcomes in the first step. Table 1 summarizes the different settings.

**Table 1:** Simulation Study: Overview

| Simulation setting | |
|---|---|
| Number of observations | 2500, 10000 |
| Covariate space dimension $p$ | $X_{i,0}, X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4}, X_{i,5}$ |
| Signal covariates in response function $\mu$ | $X_{i,0}, X_{i,1}, X_{i,2}, X_{i,3}, X_{i,5}, Z_i$ |
| Signal covariates in propensity score $P(Z_i = z\|X_i = x)$ | $X_{i,0}, X_{i,1}$ |
| Signal covariates in propensity score $P(D_i = d\|Z_i = z, X_i = x)$ | $X_{i,0}, X_{i,1}, X_{i,2}, X_{i,5}, Z_i$ |
| First step estimation | random forest |
| Second step estimation | OLS, random forest |

*Note:* This table depicts the general, DGP and estimation settings of the simulation.

### 4.4 Results

Figures 1 and 3 depict the distributions of the biases of $\hat{\theta}^{\Delta B}_{X_0}$ and $\hat{\theta}^{\Delta G}$ if the effect of interest is $\hat{\theta}^{\Delta B}_{X_0}$. It can be seen that estimating $\theta^{\Delta G}$ leads to a different result. This is because the variable $X_{i,0}$ is not balanced across the two groups of $Z_i$. On the other hand, Figures 2 and 4 depict the distributions of the biases of $\hat{\theta}^{\Delta B}_{X_2}$ and $\hat{\theta}^{\Delta G}$ for the case that the effect of interest is $\theta^{\Delta B}_{X_2}$. We see that estimating $\hat{\theta}^{\Delta G}$ leads to the same result as estimating $\hat{\theta}^{\Delta B}_{X_2}$ because the covariate $X_{i,2}$ is already balanced. Hence, if the covariate(s) is (are) not balanced, it is important to differentiate between the two effects. Due to the $\sqrt{N}$ convergence of the estimator, increasing the sample size by the factor four, should decrease the standard error by half. By comparing Figures 1 and 2 to Figures 3 and 4 we see that increasing the sample size from N$= 2,500$ to N$= 10,000$ leads to a reduction of the standard error by 50 %. Hence, this indicates a $\sqrt{N}$ convergence rate of the estimator already in finite samples.

Figures 5 and 7 depict the distributions of the biases of $\hat{\theta}^{\Delta B}_{X_0}$ when estimating the heterogeneity

[9]Simulation noise is negatively depending on the number of replications and positively on the variance of the estimator. Since the variance is doubled when the sample size is halved, we make the number of replications proportional to the sample size.

**Figure 1:** $\theta^{\Delta B}_{X_0}$ versus $\theta^{\Delta G}$ (N = 2,500)

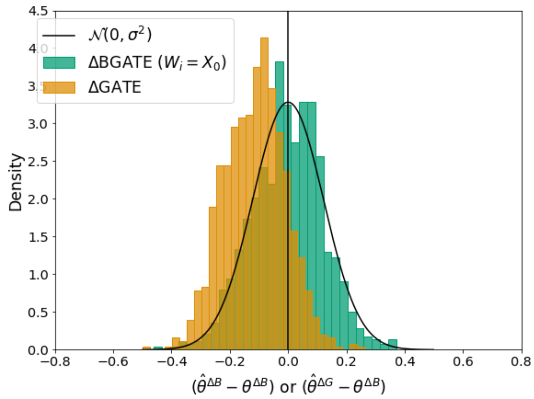



**Figure 2:** $\theta^{\Delta B}_{X_2}$ versus $\theta^{\Delta G}$ (N= 2,500)

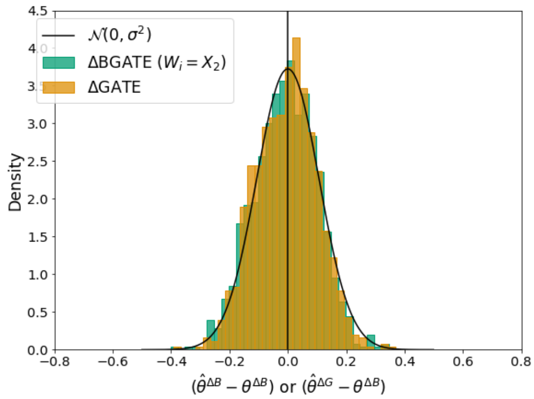



**Figure 3:** $\theta^{\Delta B}_{X_0}$ versus $\theta^{\Delta G}$ (N = 10,000)

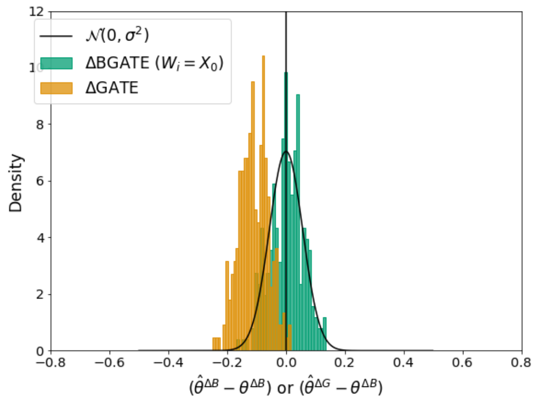
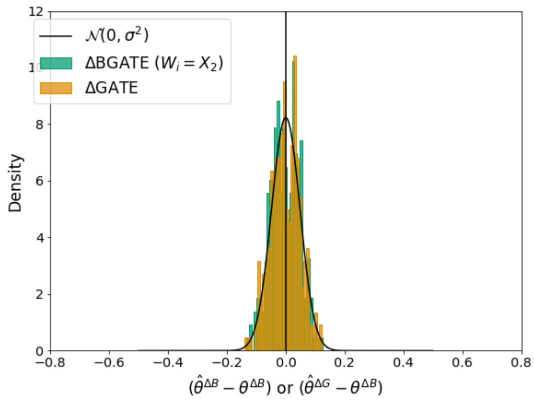



**Figure 4:** $\theta^{\Delta B}_{X_2}$ versus $\theta^{\Delta G}$ (N = 10,000)




*Note:* These figures show the distribution of the bias of $\hat{\theta}^{\Delta B}$ vs. $\hat{\theta}^{\Delta G}$ and the normal distribution with the standard deviation ($\sigma$) of $\hat{\theta}^{\Delta B}$. Column (1) shows results for a covariate that is not balanced across the groups of interest. Column (2) shows results for a covariate that is already balanced across the groups of interest. Row (1) shows results for N = 2,500 and 1,000 replications. Row (2) shows results for N = 10,000 and 250 replications. The figures are created using the results of the DGP with non-linear heterogeneity. All nuisance functions are estimated by using random forests.

with OLS despite having non-linear heterogeneous treatment effects. As expected, the results are biased. Figures 6 and 8 show the distributions of the biases of $\hat{\theta}^{\Delta B}_{X_0}$ when estimating the heterogeneous treatment effects with a random forest. Using a random forest in the second estimation step leads to better results. Again, as the comparison between the Figure 6 and 8 depict, quadrupling the sample size leads to a 50% decrease of the standard error. Hence, this indicates a $\sqrt{N}$ convergence rate of the estimator already in finite samples.

The detailed results of the simulations for all DGPs, all estimators and all sample sizes with several performance measures can be found in Appendix C.2.2.

**Figure 5:** $\theta^{\Delta B}_{X_0}$ with OLS (N = 2,500)

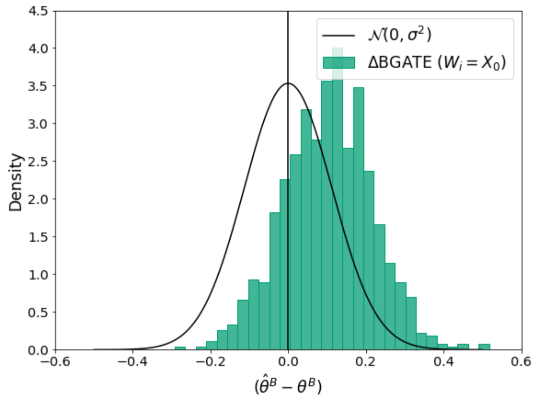



**Figure 6:** $\theta^{\Delta B}_{X_0}$ with RF (N = 2,500)

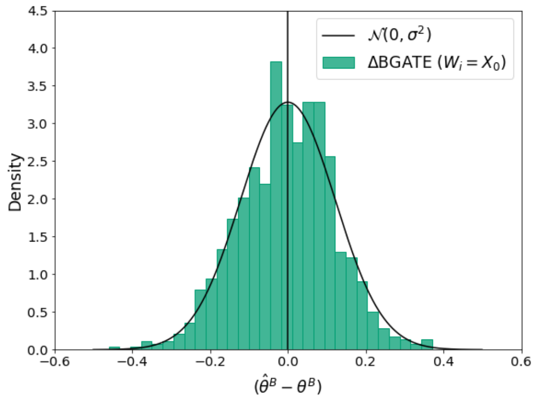



**Figure 7:** $\theta^{\Delta B}_{X_0}$ with OLS (N = 10,000)

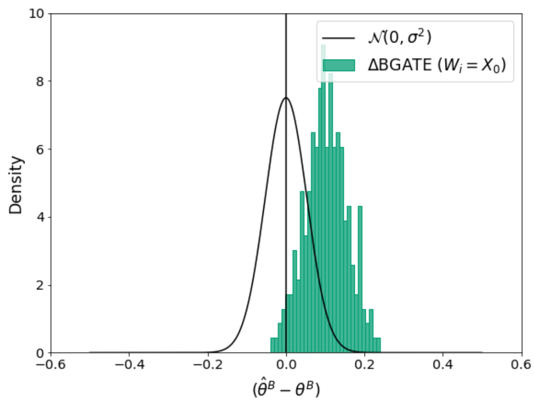
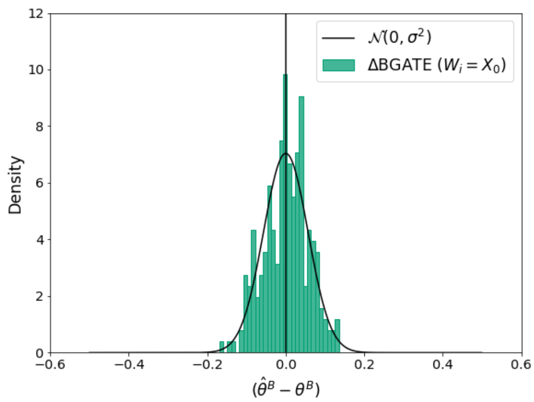



**Figure 8:** $\theta^{\Delta B}_{X_0}$ with RF (N = 10,000)




*Note:* These figures show the distribution of the bias of $\hat{\theta}^{\Delta B}_{X_0}$ and the normal distribution with the standard deviation ($\sigma$) of $\hat{\theta}^{\Delta B}$. Column (1) shows results for using OLS in the second estimation step. Column (2) shows results for using DML with a random forest in the second estimation step. Row (1) shows results for $N = 2,500$ and $1,000$ replications. Row (2) shows results for $N = 10,000$ and 250 replications. The figures are created using the results of the DGP with non-linear heterogeneity. All nuisance functions in the first estimation step are estimated by using random forests.

# 5 Illustrative Empirical Example

## 5.1 Data

To explore the method in an empirical example, we use the dataset from Lechner, Knaus, Huber, Frölich, Behncke, Mellace & Strittmatter (2020) on the effect of job search programs on the employment status of Swiss unemployed in 2003.[10] The dataset contains information on whether an individual participated in a job search program ($d = 1$) or not ($d = 0$) and an outcome variable ($Y_i$) that denotes the number of months employed in the first six months after the start of the program. The effect is claimed to be identified in an unconfoundedness setting. Hence, several covariates on the socio-economic background and the labor market history ($X_i$) of the individuals are included. The dataset consists of 91,339 unemployed individuals. Knaus, Lechner,

[10] For an extensive description of the dataset, see Knaus, Lechner, & Strittmatter (2022).

& Strittmatter (2022) find effect heterogeneity in the first six months after the start of the program. The heterogeneity relates to labor market characteristics and nationality. Individuals with disadvantaged labor market characteristics benefit more from the programs. Similarly, foreigners benefit more. These heterogeneous effects can be explained by the indirect costs of the programs (due to not searching intensively for a job during a program and therefore needing more time to find a job), which are lower for more disadvantaged and foreign individuals. We use this dataset to illustrate the proposed method and check whether these heterogeneities are due to the variables identified by the authors or whether other underlying variables confound them.

### 5.2 Summary Statistics

For the sake of brevity, we only consider the effect heterogeneity concerning nationality. Thus, we compare covariates for Swiss versus non-Swiss individuals, and treated versus non-treated individuals which helps us understand which variables might account for the variation in treatment effects of the two groups. Table 2 shows descriptive statistics for Swiss and non-Swiss individuals and treated and non-treated individuals for a selective sample of other covariates. The descriptives for all covariates can be found in Table D.1 in Appendix D.1. For descriptive statistics only by treatment status, see Table D.2 and for descriptive statistics only by moderator status, see Table D.3 in Appendix D.1. There are some differences in covariates related to the previous labor market history, namely in past income, previous job and qualifications and number of unemployment spells in the last two years. Furthermore, some socio-demographic characteristics, such as being married, also differ. Finally, as expected, there are more foreigners that have a different mother tongue than German, French, Italian or Raeto-Romansh, than Swiss individuals. Other variables, such as age or gender, are already well balanced, so balancing these covariates should not change the effect too much. However, we still include age and gender because it could happen that after balancing some (unbalanced) covariates, they are no longer balanced since they might correlate with the newly balanced covariates.

**Table 2:** Empirical analysis: Descriptive statistics for the treatment and moderator variable

| | **Treated Non-Swiss** | **Treated Swiss** | **Non-treated Non-Swiss** | **Non-treated Swiss** |
|---|---|---|---|---|
| Variable | Mean | Mean | Mean | Mean |
| Age | 35.62 | 38.08 | 36.37 | 36.50 |
| Female | 0.40 | 0.47 | 0.41 | 0.46 |
| Married | 0.67 | 0.36 | 0.70 | 0.35 |
| Mother tongue not Swiss language | 0.65 | 0.10 | 0.66 | 0.10 |
| Past annual income | 41704 | 47899 | 38226 | 43865 |
| Previous job: manager | 0.05 | 0.09 | 0.04 | 0.09 |
| Previous job: primary sector | 0.07 | 0.05 | 0.13 | 0.08 |
| Previous job: secondary sector | 0.18 | 0.15 | 0.13 | 0.13 |
| Previous job: tertiary sector | 0.54 | 0.67 | 0.48 | 0.64 |
| Previous job: missing sector | 0.21 | 0.13 | 0.26 | 0.15 |
| Previous job: self-employed | 0.00 | 0.00 | 0.01 | 0.01 |
| Previous job: skilled worker | 0.47 | 0.72 | 0.45 | 0.70 |
| Previous job: unskilled worker | 0.47 | 0.16 | 0.48 | 0.17 |
| Qualification: some degree | 0.35 | 0.73 | 0.32 | 0.73 |
| Qualification: semiskilled | 0.18 | 0.13 | 0.21 | 0.13 |
| Qualification: unskilled | 0.40 | 0.13 | 0.40 | 0.12 |
| Qualification: skilled without degree | 0.07 | 0.02 | 0.08 | 0.03 |
| No of unemployment spells last 2 years | 0.54 | 0.35 | 0.78 | 0.50 |
| Number of observations | 4,438 | 8,607 | 30,417 | 47,877 |

*Note:* This table shows the mean of some covariates included in the analysis. Column (1) and (2) show it for treated individuals, column (3) and (4) for non-treated individuals. Column (1) and (3) show it for non-Swiss individuals, column (2) and (4) for Swiss individuals.

### 5.3 Empirical Results

Table 3 shows the effects considered in the analysis. These effects include the ΔGATE, a ΔBGATE with already balanced covariates, such as age and gender, a ΔBGATE with additionally adding marital status, an extended ΔBGATE balancing additionally unbalanced covariates like past annual income, previous job variables, and qualification variables. Then, we further add the mother tongue variable. Finally, the analysis considers a ΔBGATE that balances all covariates included in the study.

Table 4 depicts the results for the different effects. As a reference point, the average treatment effect in the lock-in period[11] is $\hat{\theta} = -0.841$ (0.011). $\hat{\theta}^{\Delta G}$ shows that the difference in treatment effect between Swiss and non-Swiss individuals is significant. Hence, it seems the program works better for foreigners. However, as pointed out above, the interpretation is not straightforward because the two groups have unbalanced covariates. After balancing already balanced socio-demographic characteristics like age and gender, the coefficient remains relatively stable. When

[11] The first six months are called "lock-in" period.

**Table 3:** Empirical analysis: Effects of interest

| Effect | Variables used for balancing in the ΔBGATE |
|---|---|
| $\theta^{\Delta G}$ | none |
| $\theta_1^{\Delta B}$ | age, female |
| $\theta_2^{\Delta B}$ | + married |
| $\theta_3^{\Delta B}$ | + past income and unemployment spells, previous job variables, qualification variables |
| $\theta_4^{\Delta B}$ | + mother tongue not Swiss language |
| $\theta_5^{\Delta B}$ | all covariates included in the analysis |

*Note:* This table shows the effects of interest for the empirical analysis.

marital status is further balanced, there is a notable reduction in the coefficient. Subsequently, balancing the labor market history, including covariates like past income and previous job details, further reduces the observed difference. As anticipated, additionally, balancing mother tongue results in a minimal difference between Swiss and non-Swiss individuals of only 0.05. Balancing all covariates included in the analysis reduces the difference to zero. Hence, it becomes evident that marital status, mother tongue disparities and differences in the labor market history significantly contribute to the variance in treatment effects between Swiss and non-Swiss individuals.

These results show that researchers must carefully interpret group average treatment effects. For example, the different effect for foreigners compared to Swiss individuals is likely not caused by nationality but probably comes from different characteristics of foreigners compared to Swiss individuals.

**Table 4:** Empirical analysis: Differences of ΔBGATEs for Swiss nationals versus foreigners.

| Effect | Coef | Std. Error | P-value |
|---|---|---|---|
| $\theta^{\Delta G}$ | 0.256 | 0.037 | 0.000 |
| $\theta_1^{\Delta B}$ | 0.222 | 0.032 | 0.000 |
| $\theta_2^{\Delta B}$ | 0.188 | 0.034 | 0.000 |
| $\theta_3^{\Delta B}$ | 0.117 | 0.039 | 0.002 |
| $\theta_4^{\Delta B}$ | 0.051 | 0.040 | 0.209 |
| $\theta_5^{\Delta B}$ | -0.008 | 0.045 | 0.864 |

*Note:* This table shows the results of the empirical example ($N = 91,339$).

# 6 Causal Moderation

## 6.1 Effect of Interest

Suppose we want to interpret the difference in treatment effects between two groups causally. Then a different but closely related parameter is needed, namely the *causal balanced group average* *treatment effect ($\Delta$CBGATE).*

To define the $\Delta$CBGATE we first introduce potential outcomes that additionally depend on the moderator $Z_i$, i.e. $(Y_i^{0,0}, Y_i^{1,0}, Y_i^{0,1}, Y_i^{1,1})$.

The new parameter of interest is defined as follows:

$$\theta^{\Delta C} = \mathrm{E}\left[\left(Y_i^{1,1} - Y_i^{0,1}\right) - \left(Y_i^{1,0} - Y_i^{0,0}\right)\right] = \mathrm{E}\left[\tau(X_i, 1) - \tau(X_i, 0)\right]$$

The $\Delta$CBGATE represents the difference between the two groups, balancing the distribution of all the other covariates.[12] To identify this effect, we have to introduce new identifying assumptions.[13] The conditions to obtain a causally interpretable $\Delta$CBGATE may be challenging in many applications.

## 6.2 Identifiying Assumptions

In addition to the usual identifying assumptions stated in Section 2.2, we apply the unconfoundedness setting to the moderator variable to be able to interpret the $\Delta$CBGATE causally. Hence, the following assumptions have to be changed or added in comparison to the assumptions for the $\Delta$BGATE in Section 2.2.

**Assumption 10.** (CIA for moderator)

$$(Y_i^{1,1}, Y_i^{0,1}, Y_i^{1,0}, Y_i^{0,0}) \perp Z_i | X_i = x \quad \forall x \in \mathcal{X}$$

The CIA assumption for a moderator variable states that the potential outcomes are independent of the effect moderator $(Z_i)$, conditionally on confounding variables.

**Assumption 11.** (CS for moderator)

$$0 < P(Z_i = z | X_i = x) < 1, \quad \forall z \in \{0, 1\}, \forall x \in \mathcal{X}$$

For any given values of $X_i$ it must be possible to observe each moderator variable status $z \in \{0, 1\}$.

[12]It may not be necessary to include all available covariates to obtain causality. If the assumptions in the next section hold, then it follows that the $\Delta$CBGATE is fully balanced and other covariates do not influence this difference.

[13]In a setting with a randomized treatment and a randomized moderator variable, the $\Delta$CBGATE is identified without further assumptions.

**Assumption 12.** (Exogenity of confounders and moderator)

$$Z_i^0 = Z_i^1$$
$$X_i^{0,0} = X_i^{0,1}$$

*where $Z_d$ are potential moderators that depend on the treatment and $X_i^{d,z}$ are potential confounders that depend on the treatment and the moderator.*

New about the exogeneity assumption is that the moderator must not influence the confounders in a way related to the outcome variable. This assumption is non-standard and might be hard to fulfil in some applications. It often happens that a moderator variable such as gender influences other covariates, violating this assumption.

**Assumption 13.** (Stable Unit Treatment Value Assumption (SUTVA))

$$Y_i = \sum_{z=0}^{1} \sum_{d=0}^{1} I(d, z) Y_i^{d,z}$$

*where $I(d, z) = \mathbb{1}(D_i = d \wedge Z_i = z)$ is an indicator function which takes the value one if $D_i = d$ and $Z_i = z$ and zero otherwise.*

SUTVA now additionally requires that there are no unrepresented moderators in the population of interest (everyone is assigned to a moderation group) and that there are no relevant interactions between groups, meaning that the assignment of individual $i$ to one group does not influence the outcome of individual $j$.

**Lemma 2.** *Under Assumptions 10 to 13 the parameter $\theta^{\Delta C} = \mathrm{E}\left[\left(Y_i^{1,1} - Y_i^{0,1}\right) - \left(Y_i^{1,0} - Y_i^{0,0}\right)\right]$ is identified as $\mathrm{E}\left[\mu_1(1, X_i) - \mu_0(1, X_i) - \mu_1(0, X_i) + \mu_0(0, X_i)\right]$ with $\mu_d(z, x) = \mathrm{E}[Y_i | D_i = d, Z_i = z, X_i = x]$.*

For the proof of Lemma 2, see Appendix A.3.1. The $\Delta CBGATE$ is the same effect as a fully balanced $\Delta BGATE$ for which Assumptions 10 to 13 hold.

## 6.3 Estimation and Inference

To obtain an efficient and flexible estimator, we use DML. As shown in detail in Appendix A.3.4, the estimated Neyman-orthogonal score based on the efficient influence function for the

ΔCBGATE is given by:

$$\begin{aligned}\hat{\phi}^{\Delta C}(h;\theta^{\Delta C},\hat{\eta}) = \hat{\mu}_1(1,x) - \hat{\mu}_0(1,x) - \hat{\mu}_1(0,x) + \hat{\mu}_0(0,x) \\ + \frac{dz(y-\hat{\mu}_1(1,x))}{\hat{\omega}_{1,1}(x)} - \frac{(1-d)z(y-\hat{\mu}_0(1,x))}{\hat{\omega}_{0,1}(x)} \\ - \frac{d(1-z)(y-\hat{\mu}_1(0,x))}{\hat{\omega}_{1,0}(x)} + \frac{(1-d)(1-z)(y-\hat{\mu}_0(0,x))}{\hat{\omega}_{0,0}(x)} - \theta^{\Delta C}\end{aligned}$$

with $\hat{\omega}_{d,z}(x) = \hat{P}(D_i = d, Z_i = z|X_i = x)$, $\hat{\mu}_d(z,x) = \hat{\mathrm{E}}[Y_i|D_i = d, Z_i = z, X_i = x]$ and the nuisance parameters $\hat{\eta} = (\hat{\mu}_d(z,x), \hat{\omega}_{d,z}(x))$. Using this double robust moment condition, the ΔCBGATE is estimated at a convergence rate of $\sqrt{N}$, even if the nuisance parameters are estimated with slower rates (Smucler, Rotnitzky, & Robins, 2019). However, they must converge such that the product of the convergence rates of the outcome regression score and the propensity score estimator is faster than or equal to $\sqrt{N}$. Many common machine learning algorithms are known to converge at a rate faster or equal to $N^{1/4}$ but slower than $\sqrt{N}$ (Chernozhukov et al., 2018). Furthermore, as before, we use cross-fitting with K-folds. An estimator based on these elements is $\sqrt{N}$-consistent, asymptotically normal, and asymptotically efficient (Kennedy, 2022). Hence, the variance of $\hat{\theta}^{\Delta C}$ can be defined as

$$\begin{aligned}\mathrm{Var}(\hat{\theta}^{\Delta C}) &= \mathrm{Var}(\phi^{\Delta C}(H_i;\theta^{\Delta C},\eta)) \\ &= \mathrm{E}[\phi^{\Delta C}(H_i;\theta^{\Delta C},\eta)^2] - \underbrace{\mathrm{E}[\phi^{\Delta C}(H_i;\theta^{\Delta C},\eta)]^2}_{=0} \\ &= \mathrm{E}[\phi^{\Delta C}(H_i;\theta^{\Delta C},\eta)^2]\end{aligned}$$

and can be estimated as follows:

$$\widehat{\mathrm{Var}}(\hat{\theta}^{\Delta C}) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in S_k}[\hat{\phi}^{\Delta C}(H_i;\theta^{\Delta C},\hat{\eta})]^2$$

### 6.4 Implementation

Concerning the practical implementation, any machine learner with the required convergence rates can be used to estimate the nuisance parameters. Note that since $P(D_i = d, Z_i = z|X_i = x)$ can be rewritten as $P(D_i = d|X_i = x, Z_i = z) \cdot P(Z_i = z|X_i = x)$, there are two basic versions of the estimator. One version is based on directly estimating $P(D_i = d, Z_i = z|X_i = x)$. In contrast, the second version is based on estimating $P(D_i = d|X_i = x, Z_i = z)$ and $P(Z_i = z|X_i = x)$

separately in the full sample and subsequently obtaining the estimate of $P(D_i = d, Z_i = z|X_i = x)$ as the product of these two estimates. For the proof that the second version of the estimator is also $\sqrt{N}$-consistent and asymptotically normal, see Appendix A.3.5. Same as for the ΔBGATE, we normalize the weights (e.g.,$\frac{dz}{\omega_{1,1}(x)}$) to ensure that they do not have much more weight than the outcome regression (see Algorithm 2 in Appendix B). The proposed Algorithm 4 for the ΔCBGATE is also summarized in Appendix B. A small simulation study presented in Appendix C.3 shows the finite sample properties of the estimator and its difference to the ΔGATE.

# 7 Conclusion

This paper presents a novel approach for analyzing and interpreting treatment heterogeneity in an unconfoundedness setting. We introduce a parameter called ΔBGATE for measuring the difference in treatment effects between different groups while accounting for variations in covariates. Moreover, this paper proposes an estimator based on DML for discrete treatments and discrete moderators, demonstrating its consistency and asymptotic normality under standard conditions. A simulation study shows that the estimation strategy seems to have good finite sample properties and that estimating a ΔGATE may lead to substantially different results than estimating a ΔBGATE if the covariates are not balanced. An empirical example illustrates the proposed estimand and underlines that seeming causal heterogeneity may be caused by an underlying different distribution of other covariates. Moreover, by incorporating additional assumptions, the paper introduces the ΔCBGATE, enabling a causal interpretation of the differences in treatment effects. The proposed new parameters allow a more informative interpretation of causal heterogeneity and thus a better understanding of the differential impact of decisions. Future research could extend the estimation approach to continuous treatments and moderators. Furthermore, this paper shows identification in an unconfoundedness setting. Extending it to an instrumental variable setting for the treatment, the moderator, or both, would be interesting. More extensive simulation studies will lead to a more comprehensive picture of the finite sample properties of the proposed estimators.

## References


Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., & Spindler, M. (2024). Hyperparameter tuning for causal inference with double machine learning: A simulation study. *arXiv preprint arXiv:2402.04674*.

Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non-randomized moderators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 65–86.

Bansak, K., Bechtel, M. M., & Margalit, Y. (2021). Why austerity? The mass politics of a contested policy. *American Political Science Review*, *115*(2), 486–505.

Bansak, K., & Nowacki, T. (2022). Effect heterogeneity and causal attribution in regression discontinuity designs. *SocArXiv Papers*. Retrieved from `https://doi.org/10.31235/osf.io/vj34m`

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli Society for Mathematical Statistics and Probability*, *19*(2), 521–547.

Blackwell, M., & Olson, M. P. (2022). Reducing model misspecification and bias in the estimation of interactions. *Political Analysis*, *30*(4), 495–514.

Busso, M., DiNardo, J., & McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, *96*(5), 885–897.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68.

Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.

Clarke, P., & Polselli, A. (2023). Double machine learning for static panel models with fixed effects. *arXiv preprint arXiv:2312.08174*.

Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics*, *80*, 102306.

Davis, J. M., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, *107*(5), 546–550.

Dearing, E., & Hamilton, L. C. (2006). Contemporary advances and classic advice for analyzing mediating and moderating variables. *Monographs of the Society for Research in Child Development*, *71*(3), 88–104.

Di Francesco, R. (2023). Aggregation trees. *SSRN*.

Fairchild, A. J., & McQuillin, S. D. (2010). Evaluating mediation and moderation effects in school psychology: A presentation of methods and review of current practice. *Journal of School Psychology*, *48*(1), 53–84.

Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, *40*(1), 313–327.

Farbmacher, H., Huber, M., Lafférs, L., Langen, H., & Spindler, M. (2022). Causal mediation analysis with double machine learning. *The Econometrics Journal*, *25*(2), 277–300.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181–213.

Foster, D. J., & Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, *51*(3), 879–908.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 115.

Gogineni, A., Alsup, R., & Gillespie, D. F. (1995). Mediation and moderation in social work research. *Social Work Research*, *19*(1), 57–63.

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

Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304.

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, *175*(1), 1–21.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *79*(4), 1229–1245.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, *24*(1), 134–161.

Knaus, M. C., Lechner, M., & Strittmatter, A. (2022). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, *57*(2), 597–636.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156–4165.

Lechner, M., Knaus, M. C., Huber, M., Frölich, M., Behncke, S., Mellace, G., & Strittmatter, A. (2020). Swiss Active Labor Market Policy Evaluation [Dataset]. *Distributed by FORS, Lausanne*. Retrieved from `https://doi.org/10.23662/FORS-DS-1203-1`

Lechner, M., & Mareckova, J. (2022). Modified causal forest. *arXiv preprint arXiv:2209.03744*.

Luo, Y., Spindler, M., & Kück, J. (2016). High-dimensional l_2 boosting: Rate of convergence. *arXiv preprint arXiv:1602.08927*.

Marsh, H. W., Hau, K.-T., Wen, Z., Nagengast, B., & Morin, A. J. (2013). Moderation. In *The Oxford handbook of quantitative methods: Statistical analysis* (pp. 361–386). Oxford University Press.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Rambachan, A., Coston, A., & Kennedy, E. (2022). Counterfactual risk assessments under unmeasured confounding. *arXiv preprint arXiv:2212.09844*.

Rose, B. M., Holmbeck, G. N., Coakley, R. M., & Franks, E. A. (2004). Mediator and moderator effects in developmental and behavioral pediatric research. *Journal of Developmental & Behavioral Pediatrics*, *25*(1), 58–67.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, *219*(1), 101–122.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Smucler, E., Rotnitzky, A., & Robins, J. M. (2019). A unifying approach for doubly-robust $\ell_1$ regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.

Syrgkanis, V., & Zampetakis, M. (2020). Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory* (pp. 3453–3454).

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517–1532.

Wager, S. (2020). *STATS 361: Causal Inference.* Retrieved from `https://web.stanford.edu/~swager/stats361.pdf`

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wager, S., & Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Zimmert, M. (2018). Efficient difference-in-differences estimation with high-dimensional common trend confounding. *arXiv preprint arXiv:1809.01643*.

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

# A Appendix: Theory

## A.1 Notation

The model considered here is more general than in the main body of the paper; let the treatment be $d \in \{0, 1, \ldots, m\}$ and the moderator be $z \in \{0, 1, 2, \ldots, v\}$ variables which are discrete. To identify the effects, we rely on the usual potential outcomes by treatment $(Y_i^0, Y_i^1, \ldots, Y_i^m)$. Furthermore, we have an indicator function $I(d) = \mathbb{1}(D_i = d)$, which takes the value one if $D_i = d$ and zero otherwise, and an indicator function $I(z) = \mathbb{1}(Z_i = z)$ which takes the value one if $Z_i = z$ and zero otherwise. The analysis examines pairwise comparisons between two treatments, denoted as $m$ and $l$, and two moderator groups represented as $u$ and $v$.

## A.2 ΔBGATE

### A.2.1 Indentification based on the Outcome Regression

In this subsection, the identification of the BGATE with Assumptions 1 to 4 stated in Section 2.2 is shown. Due to the linearity in expectations assumption, the identification for the ΔBGATE directly follows. Please recall that $\mu_d(z, x) = \mathrm{E}[Y_i|D_i = d, Z_i = z, X_i = x]$.

$$
\begin{aligned}
&\mathrm{E}\Big[\mathrm{E}[Y_i^l - Y_i^m \mid Z_i = z, W_i]\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\left[\mathrm{E}\left[Y_i^l - Y_i^m \mid X_i, Z_i\right] \mid Z_i = z, W_i\right]\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\left[\mathrm{E}[Y_i^l \mid X_i, Z_i] - \mathrm{E}[Y_i^m \mid X_i, Z_i] \mid Z_i = z, W_i\right]\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\left[\mathrm{E}[Y_i^l \mid D_i = l, X_i, Z_i] - \mathrm{E}[Y_i^m \mid D_i = m, X_i, Z_i] \mid Z_i = z, W_i\right]\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\left[\mathrm{E}[Y_i \mid D_i = l, X_i, Z_i] - \mathrm{E}[Y_i \mid D_i = m, X_i, Z_i] \mid Z_i = z, W_i\right]\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\left[\mu_l(Z_i, X_i) - \mu_m(Z_i, X_i) \mid Z_i = z, W_i\right]\Big]
\end{aligned}
$$

The first equality is derived from the law of iterated expectations, while the second equality follows from the linearity in expectations. The third equality is based on Assumption 1, and the fourth is derived from the law of total expectation.

### A.2.2 Identification based on the Doubly Robust Score

The BGATE can also be identified with the doubly robust score function (see Section 3 in the main body of the text).[14] Please recall that $\pi_d(z, x) = P(D_i = d|Z_i = z, X_i = x)$ and $\lambda_z(w) =$

[14] Again, due to the linearity in expectations, this directly implies that the ΔBGATE can be identified.

$P(Z_i = z|W_i = w)$. For better readability $\mu_l(z,x) - \mu_m(z,x) + \frac{I(l)(y-\mu_l(z,x))}{\pi_l(z,x)} - \frac{I(m)(y-\mu_m(z,x))}{\pi_m(z,x)}$ is denoted as $\delta_{l,m}(h)$ and $\mathrm{E}[\delta_{l,m}(H_i)|Z_i = z, W_i = w]$ as $g_{l,m,z}(w)$.

$$
\begin{aligned}
&\mathrm{E}[\mathrm{E}[Y_i^l - Y_i^m|Z_i = z, W_i]] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[g_{l,m,z}(W_i) + \frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i\right]\right] \\
&\quad = \mathrm{E}\left[g_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i\right]\right] \\
&\quad = \mathrm{E}\left[g_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i, Z = z\right]\lambda_z(W_i)\right] \\
&\quad = \mathrm{E}\Big[g_{l,m,z}(W_i) + \mathrm{E}[\delta_{l,m}(H_i) - g_{l,m,z}(W_i)|Z_i = z, W_i]\Big] = \mathrm{E}[g_{l,m,z}(W_i)]
\end{aligned}
$$

The first equality follows from the law of iterated expectations, and the third from the law of total expectation. Hence, the BGATE can be identified if the nuisance functions are correctly specified. Due to the double robustness property, the BGATE is also identified if only the outcome regression or the propensity score is correctly specified.

*Correctly specified propensity score $\lambda_z(w)$ and wrongly specified outcome regression $\bar{g}_{l,m,z}(w)$:*

$$
\begin{aligned}
&\mathrm{E}[\mathrm{E}[Y_i^l - Y_i^m|Z_i = z, W_i]] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[\bar{g}_{l,m,z}(W_i) + \frac{I(z)(\delta_{l,m}(H_i) - \bar{g}_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i\right]\right] \\
&\quad = \mathrm{E}\left[\bar{g}_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - \bar{g}_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i\right]\right] \\
&\quad = \mathrm{E}\left[\bar{g}_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - \bar{g}_{l,m,z}(W_i))}{\lambda_z(W_i)}\middle|W_i, Z_i = z\right]\lambda_z(W_i)\right] \\
&\quad = \mathrm{E}\Big[\bar{g}_{l,m,z}(W_i) + \mathrm{E}[\delta_{l,m}(H_i) - \bar{g}_{l,m,z}(W_i)|Z_i = z, W_i]\Big] \\
&\quad = \mathrm{E}\Big[\bar{g}_{l,m,z}(W_i) + g_{l,m,z}(W_i) - \bar{g}_{l,m,z}(W_i)\Big] = \mathrm{E}[g_{l,m,z}(W_i)]
\end{aligned}
$$

*Correctly specified outcome regression $g_{l,m,z}(w)$ and wrongly specified propensity score $\bar{\lambda}_z(w)$:*

$$
\begin{aligned}
&\mathrm{E}[\mathrm{E}[Y_i^l - Y_i^m | Z_i = z, W_i]] \\
&\quad = \mathrm{E}\left[\mathrm{E}\left[g_{l,m,z}(W_i) + \frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\bar{\lambda}_z(W_i)}\middle| W_i\right]\right] \\
&\quad = \mathrm{E}\left[g_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\bar{\lambda}_z(W_i)}\middle| W_i\right]\right] \\
&\quad = \mathrm{E}\left[g_{l,m,z}(W_i) + \mathrm{E}\left[\frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\bar{\lambda}_z(W_i)}\middle| W_i, Z_i = z\right]\lambda_z(W_i)\right] \\
&\quad = \mathrm{E}\left[g_{l,m,z}(W_i) + \frac{\lambda_z(W_i)}{\bar{\lambda}_z(W_i)}\left(g_{l,m,z}(W_i) - \mathrm{E}\left[\delta_{l,m}(H_i) \mid W_i, Z_i = z\right]\right)\right] = \mathrm{E}[g_{l,m,z}(W_i)]
\end{aligned}
$$

Hence, the BGATE is identified if the outcome regression or the propensity score is correctly specified.

### A.2.3 Asymptotic Properties

The following proof is for $\theta^B_{l,m,u,v}$. However, due to the linearity in expectations, it directly follows that the proof is also valid for $\theta^{\Delta B}_{l,m,u,v}$. In a first step, let us define the following terms for easier readability:

$$
\begin{aligned}
\pi_d(z,x) &= P(D_i = d | Z_i = z, X_i = x) \\
\mu_d(z,x) &= \mathrm{E}[Y_i | D_i = d, Z_i = z, X_i = x] \\
\delta_{l,m}(h) &= \mu_l(z,x) - \mu_m(z,x) + \frac{I(l)(y - \mu_l(z,x))}{\pi_l(z,x)} - \frac{I(m)(y - \mu_m(z,x))}{\pi_m(z,x)} \\
\lambda_z(w) &= P(Z_i = z | W_i = w) \\
g_{l,m,z}(w) &= \mathrm{E}[\delta_{l,m}(H_i) | Z_i = z, W_i = w] \\
\tau_{l,m,z}(\delta_{l,m}(h), w) &= g_{l,m,z}(w) + \frac{I(z)(\delta_{l,m}(h) - g_{l,m,z}(w))}{\lambda_z(w)} \\
\theta^B_{l,m,z} &= \mathrm{E}[\tau_{l,m,z}(\delta_{l,m}(H_i), W_i)]
\end{aligned}
$$

The goal is to show that

$$
\begin{aligned}
\sqrt{N}(\hat{\theta}^B_{l,m,z} - \theta^{B\star}_{l,m,z}) &\xrightarrow{p} N(0, V^\star) \\
V^\star &= \mathrm{E}\left[\left(\tau_{l,m,z}(\delta_{l,m}(H_i), W_i) - \theta^{B\star}_{l,m,z}\right)^2\right]
\end{aligned}
$$

with $\theta^{B\star}_{l,m,z}$ being an oracle estimator of $\theta^{B}_{l,m,z}$ if all nuisance functions would be known. Then $\theta^{B\star}_{l,m,z}$ is an i.i.d. average, hence:

$$\begin{aligned}\sqrt{N}(\theta^{B\star}_{l,m,z} - \theta^{B}_{l,m,z}) &\xrightarrow{p} N(0, V^\star) \\ V^\star &= \mathrm{E}\left[\left(\tau_{l,m,z}(\delta_{l,m}(H_i), W_i) - \theta^{B\star}_{l,m,z}\right)^2\right]\end{aligned}$$

**Theorem 2.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be two half samples such that $|\mathcal{I}_1| = |\mathcal{I}_1| = \frac{N}{2}$. Furthermore, let $\mathcal{I}_{11}$, $\mathcal{I}_{12}$, $\mathcal{I}_{21}$ and $\mathcal{I}_{22}$ be four quarter samples such that $|\mathcal{I}_{11}| = |\mathcal{I}_{12}| = |\mathcal{I}_{21}| = |\mathcal{I}_{22}| = \frac{|\mathcal{I}_1|}{2} = \frac{|\mathcal{I}_2|}{2} = \frac{N}{4}$ The second split is independent conditional on the first split. Define the estimator as follows:*

$$\begin{aligned}
\hat{\theta}^{B}_{l,m,z} &= \frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} + \frac{|\mathcal{I}_{12}|}{N}\hat{\theta}^{B,\mathcal{I}_{12}}_{l,m,z} + \frac{|\mathcal{I}_{21}|}{N}\hat{\theta}^{B,\mathcal{I}_{21}}_{l,m,z} + \frac{|\mathcal{I}_{22}|}{N}\hat{\theta}^{B,\mathcal{I}_{22}}_{l,m,z} \\
\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} &= \frac{1}{|\mathcal{I}_{11}|}\sum_{\mathcal{I}_{11}} \hat{\tau}^{\mathcal{I}_{12}}_{l,m,z}(\hat{\delta}^{\mathcal{I}_2}_{l,m}(H_i), W_i) \\
\hat{\theta}^{B,\mathcal{I}_{12}}_{l,m,z} &= \frac{1}{|\mathcal{I}_{12}|}\sum_{\mathcal{I}_{12}} \hat{\tau}^{\mathcal{I}_{11}}_{l,m,z}(\hat{\delta}^{\mathcal{I}_2}_{l,m}(H_i), W_i) \\
\hat{\theta}^{B,\mathcal{I}_{21}}_{l,m,z} &= \frac{1}{|\mathcal{I}_{21}|}\sum_{\mathcal{I}_{21}} \hat{\tau}^{\mathcal{I}_{22}}_{l,m,z}(\hat{\delta}^{\mathcal{I}_1}_{l,m}(H_i), W_i) \\
\hat{\theta}^{B,\mathcal{I}_{22}}_{l,m,z} &= \frac{1}{|\mathcal{I}_{22}|}\sum_{\mathcal{I}_{22}} \hat{\tau}^{\mathcal{I}_{21}}_{l,m,z}(\hat{\delta}^{\mathcal{I}_1}_{l,m}(H_i), W_i)
\end{aligned}$$

*Then, if Assumptions 5 to 9 hold, it follows that*

$$\sqrt{N}(\hat{\theta}^{B}_{l,m,z} - \theta^{B}_{l,m,z}) \xrightarrow{d} N(0, V^\star)$$

*with*

$$V^\star = \mathrm{E}\left[\left(\tau_{l,m,z}(\delta_{l,m}(H_i), W_i) - \theta^{B\star}_{l,m,z}\right)^2\right]$$

*Proof.*

$$
\begin{aligned}
&\sqrt{N}(\hat{\theta}^B_{l,m,z} - \theta^B_{l,m,z}) \\
&\quad = \sqrt{N}(\hat{\theta}^B_{l,m,z} - \theta^{B\star}_{l,m,z} + \theta^{B\star}_{l,m,z} - \theta^B_{l,m,z}) \\
&\quad = \sqrt{N}(\hat{\theta}^B_{l,m,z} - \theta^{B\star}_{l,m,z}) + \sqrt{N}(\theta^{B\star}_{l,m,z} - \theta^B_{l,m,z}) \\
&\quad = \sqrt{N}\left(\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} + \frac{|\mathcal{I}_{12}|}{N}\hat{\theta}^{B,\mathcal{I}_{12}}_{l,m,z} + \frac{|\mathcal{I}_{21}|}{N}\hat{\theta}^{B,\mathcal{I}_{21}}_{l,m,z} + \frac{|\mathcal{I}_{22}|}{N}\hat{\theta}^{B,\mathcal{I}_{22}}_{l,m,z} - \theta^{B\star}_{l,m,z}\right) \\
&\qquad + \sqrt{N}(\theta^{B\star}_{l,m,z} - \theta^B_{l,m,z}) \\
&\quad = \sqrt{N}\Bigg(\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} + \frac{|\mathcal{I}_{12}|}{N}\hat{\theta}^{B,\mathcal{I}_{12}}_{l,m,z} + \frac{|\mathcal{I}_{21}|}{N}\hat{\theta}^{B,\mathcal{I}_{21}}_{l,m,z} + \frac{|\mathcal{I}_{22}|}{N}\hat{\theta}^{B,\mathcal{I}_{22}}_{l,m,z} \\
&\qquad - \frac{|\mathcal{I}_{11}|}{N}\theta^{B\star,\mathcal{I}_{11}}_{l,m,z} - \frac{|\mathcal{I}_{12}|}{N}\theta^{B\star,\mathcal{I}_{12}}_{l,m,z} - \frac{|\mathcal{I}_{21}|}{N}\theta^{B\star,\mathcal{I}_{21}}_{l,m,z} - \frac{|\mathcal{I}_{22}|}{N}\theta^{B\star,\mathcal{I}_{22}}_{l,m,z}\Bigg) + \sqrt{N}(\theta^{B\star}_{l,m,z} - \theta^B_{l,m,z}) \\
&\quad = \sqrt{N}\left(\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} - \frac{|\mathcal{I}_{11}|}{N}\theta^{B\star,\mathcal{I}_{11}}_{l,m,z}\right) + \sqrt{N}\left(\frac{|\mathcal{I}_{12}|}{N}\hat{\theta}^{B,\mathcal{I}_{12}}_{l,m,z} - \frac{|\mathcal{I}_{12}|}{N}\theta^{B\star,\mathcal{I}_{12}}_{l,m,z}\right) \\
&\qquad + \sqrt{N}\left(\frac{|\mathcal{I}_{21}|}{N}\hat{\theta}^{B,\mathcal{I}_{21}}_{l,m,z} - \frac{|\mathcal{I}_{21}|}{N}\theta^{B\star,\mathcal{I}_{21}}_{l,m,z}\right) + +\sqrt{N}\left(\frac{|\mathcal{I}_{22}|}{N}\hat{\theta}^{B,\mathcal{I}_{22}}_{l,m,z} - \frac{|\mathcal{I}_{22}|}{N}\theta^{B\star,\mathcal{I}_{22}}_{l,m,z}\right) \\
&\qquad + \sqrt{N}(\theta^{B\star}_{l,m,z} - \theta^B_{l,m,z})
\end{aligned}
$$

We must show that the first four terms converge to zero in probability. It is enough to show this for the first term only, as the same steps can also be directly applied to the remaining three terms.

The first summation can be decomposed as follows:

$$
\begin{aligned}
&\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} - \theta^{B\star,\mathcal{I}_{11}}_{l,m,z} \\
&= \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in I_{11}}\left(\hat{\tau}_{l,m,z}(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}, W_i) - \tau^{\star}_{l,m,z}(\delta_{l,m}(H_i), W_i)\right) \\
&= \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in \mathcal{I}_{11}}\Bigg(\hat{\mathrm{E}}[\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}|Z_i=z,W_i]^{\mathcal{I}_{12}} + \frac{I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2} - \hat{\mathrm{E}}[\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}|Z_i=z,W_i]^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} \\
&\qquad - \mathrm{E}[\delta_{l,m}(H_i)|Z_i=z,W_i] - \frac{I(z)(\delta_{l,m}(H_i) - \mathrm{E}[\delta_{l,m}(H_i)|Z_i=z,W_i])}{\lambda_z(W_i)}\Bigg) \\
&= \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in \mathcal{I}_{11}}\Bigg(\hat{\mathrm{E}}[\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}|Z_i=z,W_i]^{\mathcal{I}_{12}} + \frac{I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2} - \hat{\mathrm{E}}[\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}|Z_i=z,W_i]^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} \\
&\qquad - \hat{\mathrm{E}}[\delta_{l,m}(H_i)|Z_i=z,W_i]^{\mathcal{I}_{12}} - \frac{I(z)(\delta_{l,m}(H_i) - \hat{\mathrm{E}}[\delta_{l,m}(H_i)|Z_i=z,W_i]^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} \\
&\qquad + \hat{\mathrm{E}}[\delta_{l,m}(H_i)|Z_i=z,W_i]^{\mathcal{I}_{12}} + \frac{I(z)(\delta_{l,m}(H_i) - \hat{\mathrm{E}}[\delta_{l,m}(H_i)|Z_i=z,W_i]^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} \\
&\qquad - \mathrm{E}[\delta_{l,m}(H_i)|Z_i=z,W_i] - \frac{I(z)(\delta_{l,m}(H_i) - \mathrm{E}[\delta_{l,m}(H_i)|Z_i=z,W_i])}{\lambda_z(W_i)}\Bigg)
\end{aligned}
$$

From now on, the terms are denoted as follows: $\hat{g}_{l,m,z}(w) = \hat{\mathrm{E}}[\hat{\delta}_{l,m}(H_i)|Z_i = z, W_i = w]$, $\tilde{g}_{l,m,z}(w) = \hat{\mathrm{E}}[\delta_{l,m}(H_i)|Z_i = z, W_i = w]$ and $g_{l,m,z}(w) = \mathrm{E}[\delta_{l,m}(H_i)|Z_i = z, W_i = w]$. Next, analyze:

$$
\begin{aligned}
&\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in \mathcal{I}_{11}}\left(\hat{\tau}_{l,m,z}(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}, W_i) - \tilde{\tau}_{l,m,z}(\delta_{l,m}(H_i), W_i) + \tilde{\tau}_{l,m,z}(\delta_{l,m}(H_i), W_i) - \tau^{\star}_{l,m,z}(\delta_{l,m}(H_i), W_i)\right) \\
&\qquad = \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in \mathcal{I}_{11}}\Bigg(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} + \frac{I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2} - \hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}}_{\text{Part A}} \\
&\qquad \underbrace{- \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \frac{I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\Bigg)}_{\text{Part A}} \\
&\qquad + \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in \mathcal{I}_{11}}\left(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} + \frac{I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)}_{\text{Part B}} \\
&\qquad - \underbrace{\left(g_{l,m,z}(W_i) - \frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\right)}_{\text{Part B}}
\end{aligned}
$$

**Part A**

We start with Part A and show that $\hat{\theta}^B_{l,m,z}$ converges in probability to $\tilde{\theta}^B_{l,m,z}$ fast enough. It is possible to rewrite it in the following way:

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\Bigg(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} + \frac{I(z)(\hat{\delta}_{l,m}(H_i)^{I_2} - \hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_2}} \\
&- \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \frac{I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\Bigg) \\
&= \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left((\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\right)}_{\text{Part A.1}} \\
&+ \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\right)}_{\text{Part A.2}} \\
&+ \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2} - \hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} - \frac{1}{\lambda_z(W_i)}\right)\right)}_{\text{Part A.3}} \\
&+ \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2} - \delta_{l,m}(H_i))\right)}_{\text{Part A.4}}
\end{aligned}$$

Using the $L_2$-norm and the fact that after conditioning on $\mathcal{I}_{12}$ and $\mathcal{I}_2$ the summands become mean-zero and independent, we can show that the term $A.1$ converges in probability to zero:

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\right)^2\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\right)^2\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})^2\left(\frac{1}{\lambda_z(W_i)} - 1\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&\leqslant \frac{1}{\kappa|\mathcal{I}_{11}|}\mathrm{E}\left[(\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})^2\right] = \frac{o_p(1)}{N}
\end{aligned}$$

The second equality follows because after conditioning on $\mathcal{I}_{12}$ and $\mathcal{I}_2$ the summands become mean-zero and independent. The last inequality follows from Assumption 8. For the last equality, we need the $L_2$-norm of $\hat{g}_{l,m,z}(w)-\tilde{g}_{l,m,z}(w)$ to converge in probability and the fact that $|\mathcal{I}_{11}| = \frac{N}{4}$. Kennedy (2023) establishes the fact that

$$\begin{aligned}
&\hat{g}_{l,m,z}(w) - \tilde{g}_{l,m,z}(w) \\
&\qquad = \hat{\mathrm{E}}[\mathrm{E}[\hat{\delta}_{l,m}(H_i) - \delta_{l,m}(H_i)|Z_i = z, W_i = w, \mathcal{I}_2]|Z_i = z, W_i = w] + o_p(R_z^\star(w))
\end{aligned}$$

with

$$\begin{aligned}
R_z^\star(w)^2 = \mathrm{E}\Bigg[\Bigg(&\tilde{g}_{l,m,z}(W_i) + \frac{I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i))}{\hat{\lambda}_z(W_i)} \\
&-g_{l,m,z}(W_i) - \frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\Bigg)^2\Bigg]
\end{aligned}$$

as long as the regression estimator $\hat{\mathrm{E}}[\dots|\dots]$ is stable (Definition 1 in Kennedy (2023), Proposition B1 in Rambachan, Coston, & Kennedy (2022), Definition 1 below) with respect to distance $a$ and $a(\hat{\delta}, \delta) \xrightarrow{p} 0$.

**Definition 1.** *(Stability)*
*Suppose that the test $I_1$ and training sample $I_2$ are independent. Let:*

1. *$\hat{\delta}(h)$ be an estimate of a function $\delta(h)$ using the training data $\mathcal{I}_2$*
2. *$\hat{b}(w) = \mathrm{E}[\hat{\delta}_{l,m}(H_i) - \delta_{l,m}(H_i)|Z_i = z, W_i, \mathcal{I}_2]$ the conditional bias of the estimator $\hat{\delta}$*
3. *$\hat{\mathrm{E}}[\delta(H_i)|Z_i = z, W_i = w] = \tilde{g}_w(x)$ denote a generic regression estimator that regresses outcomes on covariates in the test sample $\mathcal{I}_1$*

*Then the regression estimator $\hat{\mathrm{E}}$ is stable with respect to distance metric $a$ at $Z_i = z$ and $W_i = w$ if:*

$$\frac{\hat{g}_z(w) - \tilde{g}_z(w) - \hat{\mathrm{E}}[\hat{b}(W_i)|Z_i = z, W_i = w]}{\sqrt{\mathrm{E}\left[\tilde{g}_z(w) - g_z(w)\right]^2}} \xrightarrow{p} 0$$

*whenever $a(\hat{\delta}, \delta) \xrightarrow{p} 0$*

Stability can be perceived as a type of stochastic equicontinuity for a nonparametric regression. They prove that linear smoothers, such as linear regressions, series regressions, nearest neighbour matching and random forests satisfy this stability condition. We will show in Part B that the oracle estimator $R_z^\star(w)$ is $o_p(1/\sqrt{N})$.

Furthermore, Kennedy (2023) shows in Theorem 2 that the bias term $\hat{b}(w)$ of an estimator regressing a doubly-robust pseudo-outcome $(\delta_{l,m}(H_i))$ on convariates $W_i$ can be expressed as:

$$\hat{b}(w) = \sum_{d=0}^{1} \frac{(\hat{\pi}_d(z,x) - \pi_d(z,x))(\hat{\mu}_d(z,x) - \mu_d(z,x))}{d\hat{\pi}_d(z,x) + (1-d)(1-\hat{\pi}_d(z,x))}$$

Therefore, as long as the estimated nuisance parameters $\hat{\pi}_d(z,x)$ and $\hat{\mu}_d(z,x)$ converge in probability to the true nuisance parameters, which is given by Assumption 6, the bias term will converge in probability to zero. In conclusion, the term A1 is $o_p(1/\sqrt{N})$.

The second term $A.2$ can again be analyzed as follows:

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\right)\right)^2\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\right)\right)^2\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\right)|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[I(z)\left(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}\right)^2\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)^2\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right] \\
&\leqslant \frac{1}{|\mathcal{I}_{11}|}(1-\kappa)\mathrm{E}\left[\left(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}\right)^2\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)^2\right] \\
&\leqslant \frac{1}{|\mathcal{I}_{11}|}\epsilon_{z1}(1-\kappa)\mathrm{E}\left[\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)^2\right] = \frac{o_p(1)}{N}
\end{aligned}$$

The third equality follows because the summands are mean-zero and independent. The last equality is true due to Assumption 6 and 8, the fact that the MSE for the inverse weights decays at the same rate as the MSE for the propensities and the fact that $|\mathcal{I}_{11}| = N/4$. Hence, the term $A.2$ is $o_p(1/\sqrt{N})$.

Similarly, this can be shown for the term $A.3$:

$$\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\hat{\lambda}_z(W_i)^{I_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\right)^2\right]$$

$$=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\right)^2\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right]$$

$$=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right]$$

$$=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[I(z)(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right]$$

$$=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[I(z)\left(\hat{\delta}_{l,m}(H_i)^{I_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}\right)^2\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\middle|\mathcal{I}_{12},\mathcal{I}_2\right]\right]$$

$$\leqslant\frac{1}{|\mathcal{I}_{11}|}(1-\kappa)\mathrm{E}\left[\left(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\hat{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}\right)^2\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\right]$$

$$\leqslant\frac{1}{|\mathcal{I}_{11}|}\epsilon_{z0}(1-\kappa)\mathrm{E}\left[\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\right]=\frac{o_p(1)}{N}$$

Again, the mean-zero and independence property of the summands is needed. The last equality is true due to Assumption 6 and 8, the fact that the MSE for the inverse weights decays at the same rate as the MSE for the propensities and the fact that $|\mathcal{I}_{11}|=N/4$. Hence, the term $A.3$ is $o_p(1/\sqrt{N})$.

Because $\delta_{l,m}(h)$ is a doubly robust score, a similar approach as for the other parts can be used again for the term $A.4$. Due to the linearity in expectations assumption, only the first part of the score function can be considered. The second part follows analogously. The term $A.4$ can be rewritten as follows:

$$\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}-\delta_{l,m}(H_i))\right)$$

$$=\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}+\frac{I(d)(Y_i-\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2})}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\mu_d(Z_i,X_i)-\frac{I(d)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)}\right)$$

$$=\underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)\right)}_{\text{Part A.4.1}}$$

$$+\underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)}_{\text{Part A.4.2}}$$

$$+\underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}I(d)\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)}_{\text{Part A.4.3}}$$

After conditioning on $I_2$, the summands used to build the term are mean-zero and independent. The squared $L_2$-norm of $A.4.1$ looks as follows:

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)\right)\right)^2\right]\\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)\right)\right)^2\Bigg|\mathcal{I}_2\right]\right]\\
&=\mathrm{E}\left[\mathrm{Var}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)\right)\right)\Bigg|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[\frac{I(z)}{\lambda_z(W_i)}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)\Bigg|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[\frac{I(z)}{\lambda_z(W_i)^2}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)^2\left(1-\frac{I(d)}{\pi_d(Z_i,X_i)}\right)^2\Bigg|\mathcal{I}_2\right]\right]\\
&\leqslant\frac{1}{\kappa^3|\mathcal{I}_{11}|}\mathrm{E}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)^2\right]=\frac{o_p(1)}{N}
\end{aligned}$$

The third line follows because the summands are mean-zero and independent. The last line follows from Assumption 5 and 6 and the fact that $|\mathcal{I}_{11}|=N/4$. Hence, Part $A.4.1$ is $o_p(1/\sqrt{N})$.

Similarly, using the squared $L_2$-norm of $A.4.2$:

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)\right)^2\right]\\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)\right)^2\middle|\mathcal{I}_2\right]\right]\\
&=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[\frac{I(z)}{\lambda_z(W_i)}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[\frac{I(z)}{\lambda_z(W_i)^2}\left(I(d)(Y_i-\mu_d(Z_i,X_i))\right)^2\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)^2\middle|\mathcal{I}_2\right]\right]\\
&\leqslant\frac{1}{\kappa^2|\mathcal{I}_{11}|}(1-\kappa)\mathrm{E}\left[\mathrm{E}\left[\left((Y_i-\mu_d(Z_i,X_i))\right)^2\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)^2\middle|\mathcal{I}_2\right]\right]\\
&\leqslant\frac{1}{\kappa^2|\mathcal{I}_{11}|}(1-\kappa)\epsilon_1\mathrm{E}\left[\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)^2\right]=\frac{o_p(1)}{N}
\end{aligned}$$

The third line follows from the fact that the summands are mean-zero and independent. The last two inequalitites follow from Assumption 5, 6 and 8 and the fact that $|\mathcal{I}_{11}|=N/4$.

Last, using the $L_1$-norm of $A.4.3$:

$$\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\frac{I(z)}{\lambda_z(W_i)}I(d)\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right)\right|\right]\\
&\quad\leqslant\frac{1}{\kappa}\mathrm{E}\left[\left|\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)\right|\right]\\
&\quad\leqslant\frac{1}{\kappa}\mathrm{E}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\left|\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right|\right)\right]\\
&\quad=\frac{1}{\kappa}\mathrm{E}\left[\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\left|\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right|\right]\\
&\quad\leqslant\frac{1}{\kappa}\mathrm{E}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)^2\right]^{1/2}\mathrm{E}\left[\left(\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}}-\frac{1}{\pi_d(Z_i,X_i)}\right)^2\right]^{1/2}=\frac{o_p(1)}{\sqrt{N}}
\end{aligned}$$

The first inequality follows from Assumption 5, the second inequality follows from Cauchy-Schwarz, the last line from Assumption 7 and the last equality from Assumption 6 and the fact that $|\mathcal{I}_{11}|=N/4$. Hence, term $A.4.3$ is $o_p(1/\sqrt{N})$.

Summing up, we have shown that Part $A$ is $o_p(1/\sqrt{N})$ as long as the product of the estimation errors decays faster than $1/\sqrt{N}$, which is given by Assumption 7.

$$\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\hat{\tau}_{l,m,z}(\hat{\delta}_{l,m}(H_i)^{\mathcal{I}_2}, W_i)^{\mathcal{I}_{11}} - \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\tilde{\tau}_{l,m,z}(\delta_{l,m}(H_i), W_i)^{\mathcal{I}_{11}} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

Hence, it follows that

$$\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} - \frac{|\mathcal{I}_{11}|}{N}\tilde{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

**Part B**

In the next step, the term $B$ is considered. It is the same proof as the usual proof for the average treatment effect (Wager, 2020) since we assume that the true pseudo-outcome $\delta_{l,m}(h)$ is known. The term can be rewritten as follows:

$$\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} + \frac{I(z)(\delta_{l,m}(H_i) - \tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}})}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} - g_{l,m,z}(W_i) - \frac{I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))}{\lambda_z(W_i)}\right)$$
$$= \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left((\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i))\left(1 - \frac{I(z)}{\lambda_z(W_i)}\right)\right)}_{\text{Part B.1}}$$
$$+ \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i) - g_{l,m,z}(W_i))\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}} - \frac{1}{\lambda_z(W_i)}\right)\right)}_{\text{Part B.2}}$$
$$+ \underbrace{\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i))\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\right)}_{\text{Part B.3}}$$

Still following Wager (2020), we can show that all three terms converge to zero in probability. For the term $B.1$, after conditioning on $\mathcal{I}_{12}$, the summands used to build the term are mean-zero and independent. Using the squared $L_2$-norm of $B.1$:

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i))\left(1-\frac{I(z)}{\lambda_z(W_i)}\right)\right)^2\right]\\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i))\left(1-\frac{I(z)}{\lambda_z(W_i)}\right)\right)^2\middle|\mathcal{I}_{12}\right]\right]\\
&=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left((\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i))\left(1-\frac{I(z)}{g_{l,m,z}(W_i)}\right)\right)\middle|\mathcal{I}_{12}\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[\left(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i)\right)\left(1-\frac{I(z)}{\lambda_z(W_i)}\right)\middle|\mathcal{I}_{12}\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[\left(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i)\right)^2\left(\frac{1}{\lambda_z(W_i)}-1\right)\middle|\mathcal{I}_{12}\right]\right]\\
&\leqslant\frac{1}{\kappa|\mathcal{I}_{11}|}\mathrm{E}\left[(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}}-g_{l,m,z}(W_i))^2\right]=\frac{o_p(1)}{N}
\end{aligned}
$$

The second equality follows because the summands are mean-zero and independent. The last line follows from Assumption 5 and 6 and the fact that $|\mathcal{I}_{11}|=N/4$. Hence, the term $B.1$ is $o_p(1/\sqrt{N})$.

Similarly, using the squared $L_2$-norm of $B.2$:

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i)-g_{l,m,z}(W_i))\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\right)^2\right]\\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i)-g_{l,m,z}(W_i))\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\right)^2\middle|\mathcal{I}_{12}\right]\right]\\
&=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\left(I(z)(\delta_{l,m}(H_i)-g_{l,m,z}(W_i))\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\right)\middle|\mathcal{I}_{12}\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{Var}\left[I(z)\left(\delta_{l,m}(H_i)-g_{l,m,z}(W_i)\right)\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)\middle|\mathcal{I}_{12}\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[\mathrm{E}\left[I(z)(\delta_{l,m}(H_i)-g_{l,m,z}(W_i))^2\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\middle|\mathcal{I}_{12}\right]\right]\\
&=\frac{1}{|\mathcal{I}_{11}|}\mathrm{E}\left[I(z)\left(\delta_{l,m}(H_i)-g_{l,m,z}(W_i)\right)^2\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\right]\\
&\leqslant\frac{1}{|\mathcal{I}_{11}|}(1-\kappa)\mathrm{E}\left[(\delta_{l,m}(H_i)-g_{l,m,z}(W_i))^2\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\right]\\
&\leqslant\frac{1}{|\mathcal{I}_{11}|}(1-\kappa)\epsilon_{z1}\mathrm{E}\left[\left(\frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}-\frac{1}{\lambda_z(W_i)}\right)^2\right]=\frac{o_p(1)}{N}
\end{aligned}
$$

Again, the second equality follows because the summands are mean-zero and independent. The last two inequalities follow from Assumption 6 and 8, the fact that the MSE for the inverse weights decays at the same rate as the MSE for the propensities and the fact that $|\mathcal{I}_{11}| = N/4$. Hence, the term $B.2$ is $o_p(1/\sqrt{N})$.

Last, using the $L_1$-norm of $B.3$:

$$\begin{aligned}
&\mathrm{E}\Bigg[\Bigg|\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}}\Bigg(I(z)(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i))\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)\Bigg)\Bigg|\Bigg] \\
&\quad\leqslant \mathrm{E}\Bigg[\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} I(z)\Big|\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i)\Big|\Big|\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\Big|\Bigg] \\
&\quad= \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} \mathrm{E}\Bigg[I(z)\Big|\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i)\Big|\Big|\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\Big|\Bigg] \\
&\quad= \mathrm{E}\Bigg[I(z)\Big|\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i)\Big|\Big|\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\Big|\Bigg] \\
&\quad\leqslant \mathrm{E}\Big[I(z)(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i))^2\Big]^{1/2} \mathrm{E}\Bigg[I(z)\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)^2\Bigg]^{1/2} \\
&\quad\leqslant \mathrm{E}\Big[(\tilde{g}_{l,m,z}(W_i)^{\mathcal{I}_{12}} - g_{l,m,z}(W_i))^2\Big]^{1/2} \mathrm{E}\Bigg[\left(\frac{1}{\lambda_z(W_i)} - \frac{1}{\hat{\lambda}_z(W_i)^{\mathcal{I}_{12}}}\right)^2\Bigg]^{1/2} = \frac{o_p(1)}{\sqrt{N}}
\end{aligned}$$

The first inequality follows from Cauchy-Schwarz, the fourth line from Assumption 7 and the last equality from Assumption 6 and the fact that $|\mathcal{I}_{11}| = N/4$. Hence, term $B.2$ is $o_p(1/\sqrt{N})$.

Hence, we have shown that Part $B$ is $o_p(1/\sqrt{N})$ as long as the product of the estimation errors decays faster than $1/\sqrt{N}$, which is given by Assumption 7. Summing up, in Part B, we have shown that:

$$\frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} \tilde{\tau}_{l,m,z}(\delta(H_i), W_i)^{\mathcal{I}_{11}} - \frac{1}{|\mathcal{I}_{11}|}\sum_{i\in\mathcal{I}_{11}} \tau^{\star}_{l,m,z}(\delta(H_i), W_i)^{\mathcal{I}_{11}} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

Hence, it follows that

$$\frac{|\mathcal{I}_{11}|}{N}\tilde{\theta}^{B,\mathcal{I}_{11}}_{l,m,z} - \frac{|\mathcal{I}_{11}|}{N}\theta^{B\star,\mathcal{I}_{11}}_{l,m,z} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

and therefore,

$$\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}} - \frac{|\mathcal{I}_{11}|}{N}\theta^{B\star,\mathcal{I}_{11}} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

Putting all the parts together, we conclude that:

$$
\begin{aligned}
&\sqrt{N}(\hat{\theta}^B - \theta^B) \\
&\quad = \underbrace{\sqrt{N}\bigg(\frac{|\mathcal{I}_{11}|}{N}\hat{\theta}^{B,\mathcal{I}_{11}} - \frac{|\mathcal{I}_{11}|}{N}\theta^{B\star,\mathcal{I}_{11}}\bigg)}_{=o_p(1)} + \underbrace{\sqrt{N}\bigg(\frac{|\mathcal{I}_{12}|}{N}\hat{\theta}^{B,\mathcal{I}_{12}} - \frac{|\mathcal{I}_{12}|}{N}\theta^{B\star,\mathcal{I}_{12}}\bigg)}_{=o_p(1)} \\
&\quad + \underbrace{\sqrt{N}\bigg(\frac{|\mathcal{I}_{21}|}{N}\hat{\theta}^{B,\mathcal{I}_{21}} - \frac{|\mathcal{I}_{21}|}{N}\theta^{B\star\mathcal{I}_{21}}\bigg)}_{=o_p(1)} + + \underbrace{\sqrt{N}\bigg(\frac{|\mathcal{I}_{22}|}{N}\hat{\theta}^{B,\mathcal{I}_{22}} - \frac{|\mathcal{I}_{22}|}{N}\theta^{B\star\mathcal{I}_{22}}\bigg)}_{=o_p(1)} \\
&\quad + \underbrace{\sqrt{N}(\theta^{B\star} - \theta^B)}_{\xrightarrow{d} N(0,V^\star)}
\end{aligned}
$$

Hence, the estimator $\hat{\theta}^B$ is $\sqrt{N}$-consistent and asymptotically normal.

## A.3 ΔCBGATE

### A.3.1 Identification based on the Outcome Regression

This subsection identifies the ΔCBGATE with the assumptions from Section 2.2. Please recall that $\mu_d(z,x) = \mathrm{E}[Y_i|D_i = d, Z_i = z, X_i = x]$. The aim is to show that we can estimate the estimand of interest $\mathrm{E}\Big[\Big(Y_i^{m,v} - Y_i^{l,v}\Big) - \Big(Y_i^{m,u} - Y_i^{l,u}\Big)\Big]$.

$$
\begin{aligned}
&\mathrm{E}[(Y_i^{m,v} - Y_i^{l,v}) - (Y_i^{m,u} - Y_i^{l,u})] \\
&\quad = \mathrm{E}\Big[\mathrm{E}\Big[\Big(Y_i^{m,v} - Y_i^{l,v}\Big) - \Big(Y_i^{m,u} - Y_i^{l,u}\Big) \mid X_i\Big]\Big] \\
&\quad = \mathrm{E}\Big[\mathrm{E}\left[Y_i^{m,v} \mid X_i\right] - \mathrm{E}\Big[Y_i^{l,v} \mid X_i\Big] - \mathrm{E}\left[Y_i^{m,u} \mid X_i\right] + \mathrm{E}\Big[Y_i^{l,u} \mid X_i\Big]\Big] \\
&\quad = \mathrm{E}\Big[\mathrm{E}\left[Y_i^{m,v} \mid Z_i = v, X_i\right] - \mathrm{E}\Big[Y_i^{l,v} \mid Z_i = v, X_i\Big] - \mathrm{E}\left[Y_i^{m,u} \mid Z_i = u, X_i\right] + \mathrm{E}\Big[Y_i^{l,u} \mid Z_i = u, X_i\Big]\Big] \\
&\quad = \mathrm{E}\Big[\mathrm{E}\left[Y_i^{m,v} \mid D_i = m, Z_i = v, X_i\right] - \mathrm{E}\Big[Y_i^{l,v} \mid D_i = l, Z_i = v, X_i\Big] \\
&\qquad - \mathrm{E}\left[Y_i^{m,u} \mid D_i = m, Z_i = u, X_i\right] + \mathrm{E}\Big[Y_i^{l,u} \mid D_i = l, Z_i = u, X_i\Big]\Big] \\
&\quad = \mathrm{E}\Big[\mathrm{E}[Y_i|D_i = m, Z_i = v, X_i] - \mathrm{E}[Y_i|D_i = l, Z_i = v, X_i] \\
&\qquad - \mathrm{E}[Y_i|D_i = m, Z_i = u, X_i] + \mathrm{E}[Y_i|D_i = l, Z_i = u, X_i]\Big] \\
&\quad = \mathrm{E}[\mu_m(v, X_i) - \mu_l(v, X_i) - \mu_m(u, X_i) + \mu_l(u, X_i)]
\end{aligned}
$$

The first equality follows from the law of iterated expectations, the second from the linearity of expectations and the third from Assumption 10. The fourth equality follows from Assumption

1, and the fifth equality follows the law of total expectation.

### A.3.2 Identification based on the Doubly Robust Score

The parameter of interest can also be identified with the doubly robust score function. Please recall that $\omega_{d,z}(x) = P(D_i = d, Z_i = z|X_i = x)$ and $I(d,z) = \mathbb{1}(D_i = d \wedge Z_i = z)$. It is enough to show that the average potential outcome $\mathrm{E}[Y_i^{d,z}]$ is identified due to the linearity of expectation property:

$$
\begin{aligned}
\mathrm{E}[Y_i^{d,z}] &= \mathrm{E}\left[\mathrm{E}\left[\mu_d(z,X_i) + \frac{I(d,z)(Y_i - \mu_d(z,X_i))}{\omega_{d,z}(X_i)}|X_i\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mu_d(z,X_i) + \frac{I(d,z)(Y_i - \mu_d(z,X_i))}{\omega_{d,z}(X_i)}|D_i = d, Z_i = z, X_i\right]\omega_{d,z}(X_i)\right] \\
&= \mathrm{E}[\mathrm{E}[\mu_d(z,X_i) + Y_i - \mu_d(z,X_i)|D_i = d, Z_i = z, X_i]] \\
&= \mathrm{E}[\mathrm{E}[Y_i|D_i = d, Z_i = z, X_i]] \\
&= \mathrm{E}[\mu_d(z,X_i)]
\end{aligned}
$$

The first equality is due to the linearity of expectations, and the second is due to the law of total expectations.

### A.3.3 Neyman Orthogonality

To be able to use machine learning algorithms to estimate the nuisance functions, the score function has to be Neyman orthogonal. This section follows closely Knaus (2022). We use the following APO (average potential outcome) building block to build the double-double robust estimator:

$$
\Gamma_{d,z}(h) = \mu_d(z,x) + \frac{I(d,z)(y - \mu_d(z,x))}{\omega_{d,z}(x)}
$$

The estimate of interest $\theta^{\Delta C} = \Delta\mathrm{CBGATE}$ can be built from the four different APO's, and this looks as follows:

$$
\theta^{\Delta C} = \mathrm{E}[\Gamma_{m,v}(H_i) - \Gamma_{l,v}(H_i) - \Gamma_{m,u}(H_i) + \Gamma_{l,u}(H_i)]
$$

Let us show that the APO is Neyman-orthogonal. The score looks the following

$$\mathrm{E}\left[\underbrace{\mu_d(z,X_i) + \frac{I(d,z)(Y_i - \mu_d(z,X_i))}{\omega_{d,z}(X_i)} - \psi_{d,z}}_{\phi(H_i;\psi_{d,z},\mu_d(z,X_i),\omega_{d,z}(X_i))}\right] = 0$$

with $\psi_{d,z} = \mathrm{E}[\Gamma_{d,z}(H_i)]$. A score $\phi(h;\psi_{d,z},\mu_d(z,x),\omega_{d,z}(x))$ is Neyman-orthogonal if its Gateaux derivative w.r.t. to the nuisance parameters is in expectation zero at the true nuisance parameters. This means the following:

$$\partial_r \mathrm{E}[\phi(H_i;\psi_{d,z},\mu_d(z,X_i)+r(\tilde{\mu}_d(z,X_i)-\mu_d(z,X_i)),\omega_{d,z}(X_i)+r(\tilde{\omega}_{d,z}(X_i)-\omega_{d,z}(X_i)) \mid X_i] \mid_{r=0} = 0.$$

To show that the APO is Neyman-orthogonal, we first have to add the perturbation to the nuisance parameters of the score:

$$\begin{aligned}
&\phi(h;\psi_{d,z},\mu + r(\tilde{\mu}_d(z,x) - \mu_d(z,x)),\omega_{d,z}(x) + r(\tilde{\omega}_{d,z}(x) - \omega_{d,z}(x))) \\
&\quad = (\mu_d(z,x) + r(\tilde{\mu}_d(z,x) - \mu_d(z,x))) + \frac{I(d,z)y - I(d,z)(\mu_d(z,x) + r(\tilde{\mu}_d(z,x) - \mu_d(z,x)))}{\omega_{d,z}(x) + r(\tilde{\omega}_{d,z}(x) - \omega_{d,z}(x))} \\
&\qquad - \psi_{d,z}
\end{aligned}$$

In a second step, the conditional expectation is taken

$$\begin{aligned}
&\mathrm{E}[\phi(H_i;\psi_{d,z},\mu_d(z,X_i) + r(\tilde{\mu}_d(z,X_i) - \mu_d(z,X_i)),\omega_{d,z}(X_i) + r(\tilde{\omega}_{d,z}(X_i) - \omega_{d,z}(X_i))) \mid X_i = x] \\
&\quad = \mathrm{E}\Bigg[(\mu_d(z,X_i) + r(\tilde{\mu}_d(z,X_i) - \mu_d(z,X_i))) \\
&\qquad + \frac{I(d,z)Y_i - I(d,z)(\mu_d(z,X_i) + r(\tilde{\mu}_d(z,X_i) - \mu_d(z,X_i)))}{\omega_{d,z}(X_i) + r(\tilde{\omega}_{d,z}(X_i) - \omega_{d,z}(X_i))} - \psi_{d,z} \Bigg| X_i = x\Bigg] \\
&\quad = (\mu_d(z,X_i) + r(\tilde{\mu}_d(z,X_i) - \mu_d(z,X_i))) + \mathrm{E}\left[\frac{I(d,z)Y_i}{\omega_{d,z}(X_i) + r(\tilde{\omega}_{d,z}(X_i) - \omega_{d,z}(X_i))}\middle| X_i = x\right] \\
&\qquad - \mathrm{E}\left[\frac{I(d,z)(\mu_d(z,X_i) + r(\tilde{\mu}_d(z,X_i) - \mu_d(z,X_i)))}{\omega_{d,z}(X_i) + r(\tilde{\omega}_{d,z}(X_i) - \omega_{d,z}(X_i))}\middle| X_i = x\right] - \psi_{d,z} \\
&\quad = (\mu_d(z,x) + r(\tilde{\mu}_d(z,x) - \mu_d(z,x))) + \frac{\mu_d(z,x)\omega_{d,z}(x)}{\omega_{d,z}(x) + r(\tilde{\omega}_{d,z}(x) - \omega_{d,z}(x))} \\
&\qquad - \frac{\omega_{d,z}(x)(\mu_d(z,x) + r(\tilde{\mu}_d(z,x) - \mu_d(z,x)))}{\omega_{d,z}(x) + r(\tilde{\omega}_{d,z}(x) - \omega_{d,z}(x))} - \psi_{d,z}
\end{aligned}$$

The third equality follows from:

$$\begin{aligned}
\mathrm{E}[I(d,z)Y_i|X_i=x] &= \mathrm{E}[I(d,z)\sum_d\sum_z I(d,z)Y_i^{d,z}|X_i=x] \\
&= \mathrm{E}[I(d,z)Y_i^{d,z}|X_i=x] \\
&= \mu_d(z,x)\omega_{d,z}(x)
\end{aligned}$$

In a third step, the derivative with respect to r is taken:

$$\begin{aligned}
&\partial_r \mathrm{E}[\phi(H_i;\psi_{d,z},\mu_d(z,X_i)+r(\tilde{\mu}_d(z,X_i)-\mu_d(z,X_i)),\omega_{d,z}(X_i)+r(\tilde{\omega}_{d,z}(X_i)-\omega_{d,z}(X_i)))\mid X_i=x] \\
&= (\tilde{\mu}_d(z,x)-\mu_d(z,x)) - \frac{\mu_d(z,x)\omega_{d,z}(x)(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{(\omega_{d,z}(x)+r(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x)))^2} \\
&- \frac{\omega_{d,z}(x)(\tilde{\mu}_d(z,x)-\mu_d(z,x))(\omega_{d,z}(x)+r(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x)))}{(\omega_{d,z}(x)+r(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x)))^2} \\
&- \frac{\omega_{d,z}(x)(\mu_d(z,x)-r(\tilde{\mu}_d(z,x)-\mu_d(z,x)))(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{(\omega_{d,z}(x)+r(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x)))^2}
\end{aligned}$$

Finally, evaluate at the true nuisance values, i.e. set r = 0:

$$\begin{aligned}
&\partial_r \mathrm{E}[\phi(H_i;\psi_{d,z},\mu_d(z,X_i)+r(\tilde{\mu}_d(z,X_i)-\mu_d(z,X_i)),\omega_{d,z}(X_i)+r(\tilde{\omega}_{d,z}(X_i)-\omega_{d,z}(X_i)))\mid X_i=x]\,|_{r=0} \\
&= (\tilde{\mu}_d(z,x)-\mu_d(z,x)) - \frac{\mu_d(z,x)\omega_{d,z}(x)(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{\omega_{d,z}(x)^2} \\
&- \frac{\omega_{d,z}(x)(\tilde{\mu}_d(z,x)-\mu_d(z,x))\omega_{d,z}(x)-\omega_{d,z}(x)\mu_d(z,x)(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{\omega_{d,z}(x)^2} \\
&= (\tilde{\mu}_d(z,x)-\mu_d(z,x)) - \frac{\mu_d(z,x)\omega_{d,z}(x)(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{\omega_{d,z}(x)^2} \\
&- \frac{\omega_{d,z}(x)^2(\tilde{\mu}_d(z,x)-\mu_d(z,x))}{\omega_{d,z}(x)^2} + \frac{\omega_{d,z}(x)\mu_d(z,x)(\tilde{\omega}_{d,z}(x)-\omega_{d,z}(x))}{\omega_{d,z}(x)^2} \\
&= 0
\end{aligned}$$

### A.3.4 Derivation of the Influence Function

The Gateaux derivative approach is used to derive the influence function, as shown in Section 3.4.2 in Kennedy (2022). Furthermore, similarly to showing that the score function is Neyman-orthogonal where we closely rely on Knaus (2022) (see Appendix A.3.3), the influence function derivation is based on one APO and then extend it as Hines, Dukes, Diaz-Ordaz, & Vansteelandt (2022) show it for the Average Treatment Effect (ATE). Hence, we derive the influence function for the functional $\psi_{d,z} = \mathrm{E}(Y_i^{d,z}) = \mathrm{E}[\mathrm{E}[Y_i|D_i=d, Z_i=z, X_i]]$.

A particular choice of a parametric submodel is used, for which the pathwise derivative is equal to the influence function.

**Definition 2.** (Definition 1 in Kennedy (2022)) *A parametric submodel is a smooth parametric model* $P_\epsilon = \{P_\epsilon : \epsilon \in \mathbb{R}\}$ *that satisfiers (i)* $P_\epsilon \subseteq P$*, and (ii)* $P_{\epsilon=0} = \mathbb{P}$.

First, we have to assume that the data is discrete. This simplifies the calculations. If the regressions functions $\mu_d(z,x) = \mathrm{E}[Y_i|D_i = d, Z_i = z, X_i = x]$ and $\omega_{d,z}(x) = P(D_i = d, Z_i = z|X_i = x)$ are well-defined outside of the discrete setup, the influence function is also well-defined (Kennedy, 2022). For the further derivation, $H_i = (D_i, Y_i, Z_i, X_i)$ and $\mathbb{P}$ is the true distribution. The simple submodel is given by $\mathbb{P}^\star_\epsilon(H_i) = (1-\epsilon)\mathbb{P}(H_i) + \epsilon\delta_h$ where $\delta_h$ is the Dirac measure at $H_i = h$. Since $H_i$ is discrete, we can work with the mass function $P^\star_\epsilon(H_i) = (1-\epsilon)P(H_i) + \epsilon\mathbb{1}(H_i = h)$. Note that for the submodel, we have

$$
\begin{aligned}
P^\star_\epsilon(Y_i|D_i, Z_i, X_i) &= \frac{P^\star_\epsilon(H_i)}{P^\star_\epsilon(D_i, Z_i, X_i)} = \frac{(1-\epsilon)P(H_i) + \epsilon\mathbb{1}(H_i = h)}{(1-\epsilon)P(D_i, Z_i, X_i) + \epsilon\mathbb{1}(D_i = d, Z_i = z, X_i = x)} \\
P^\star_\epsilon(D_i|Z_i, X_i) &= \frac{P^\star_\epsilon(D_i, Z_i, X_i)}{P^\star_\epsilon(Z_i, X_i)} = \frac{(1-\epsilon)P(D_i, Z_i, X_i) + \epsilon\mathbb{1}(D_i = d, Z_i = z, X_i = x)}{(1-\epsilon)P(Z_i, X_i) + \epsilon\mathbb{1}(Z_i = z, X_i = x)} \\
P^\star_\epsilon(Z_i|X_i) &= \frac{P^\star_\epsilon(Z_i, X_i)}{P^\star_\epsilon(X_i)} = \frac{(1-\epsilon)P(Z_i, X_i) + \epsilon\mathbb{1}(Z_i = z, X_i = x)}{(1-\epsilon)P(X_i) + \epsilon\mathbb{1}(X_i = x)} \\
P^\star_\epsilon(X_i) &= (1-\epsilon)P(X_i) + \epsilon\mathbb{1}(X_i = x)
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial}{\partial\epsilon} P^\star_\epsilon(Y_i|D_i, Z_i, X_i)\Big|_{\epsilon=0} \\
&= \frac{\mathbb{1}(H_i = h) - P(H_i)}{(1-\epsilon)P(D_i, Z_i, X_i) + \epsilon\mathbb{1}(D_i = d, Z_i = z, X_i = x)}\Bigg|_{\epsilon=0} \\
&\quad - P^\star_\epsilon(Y_i|D_i, Z_i, X_i)\frac{\mathbb{1}(D_i = d, Z_i = z, X_i = x) - P(D_i, Z_i, X_i)}{(1-\epsilon)P(D_i, Z_i, X_i) + \epsilon\mathbb{1}(D_i = d, Z_i = z, X_i = x)}\Bigg|_{\epsilon=0} \\
&= \frac{\mathbb{1}(H_i = h) - P(H_i)}{P(D_i, Z_i, X_i)} - P(Y_i|D_i, Z_i, X_i)\frac{\mathbb{1}(D_i = d, Z_i = z, X_i = x) - P(D_i, Z_i, X_i)}{P(D_i, Z_i, X_i)} \\
&= \frac{\mathbb{1}(H_i = h)}{P(D_i, Z_i, X_i)} - \underbrace{\frac{P(H_i)}{P(D_i, Z_i, X_i)}}_{= P(Y_i|D_i, Z_i, X_i)} - P(Y_i|D_i, Z_i, X_i)\frac{\mathbb{1}(D_i = d, Z_i = z, X_i = x)}{P(D_i, Z_i, X_i)} \\
&\quad + P(Y|D_i, Z_i, X_i)\frac{P(D_i, Z_i, X_i)}{P(D_i, Z_i, X_i)} \\
&= \frac{\mathbb{1}(H_i = h)}{P(D_i, Z_i, X_i)} - P(Y_i|D_i, Z_i, X_i)\frac{\mathbb{1}(D_i = d, Z_i = z, X_i = x)}{P(D_i, Z_i, X_i)} \\
&= \mathbb{1}(D_i = d, Z_i = z, X_i = x)\left[\frac{\mathbb{1}(Y_i = y) - P(Y_i|D_i, Z_i, X_i)}{P(D_i, Z_i, X_i)}\right]
\end{aligned}
$$

In the next step we evaluate the parameter on the submodel, differentiate, and set $\epsilon = 0$.

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\psi(P^\star_\epsilon)\bigg|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon}\sum_{X,Y} P^\star_\epsilon(Y_i|D_i,Z_i,X_i)P^\star_\epsilon(X_i)\bigg|_{\epsilon=0} \\
&= \sum_{X,Y} Y\left[\frac{\partial}{\partial \epsilon}P^\star_\epsilon(Y_i|D_i,Z_i,X_i)P^\star_\epsilon(X_i) + P^\star_\epsilon(Y_i|D_i,Z_i,X_i)\frac{\partial}{\partial \epsilon}P^\star_\epsilon(X)\right]\bigg|_{\epsilon=0} \\
&= \sum_{X,Y} Y_i\Bigg[\mathbb{1}(D_i = d, Z_i = z, X_i = x)\left(\frac{\mathbb{1}(Y_i = y) - P(Y_i|D_i,Z_i,X_i)}{P(D_i,Z_i,X_i)}\right)P(X_i) \\
&\quad + P(Y_i|D_i,Z_i,X_i)\left(\mathbb{1}(X_i = x) - P(X_i)\right)\Bigg] \\
&= \sum_{X,Y}\Bigg[\frac{Y_i\mathbb{1}(D_i = d, Z_i = z, X_i = x)\mathbb{1}(Y_i = y)P(X_i)}{P(D_i,Z_i,X_i)} \\
&\quad - \frac{Y_i\mathbb{1}(D_i = d, Z_i = z, X_i = x)P(Y_i|D_i,Z_i,X_i)P(X_i)}{P(D_i,Z_i,X_i)} \\
&\quad + Y P(Y_i|D_i,Z_i,X_i)\mathbb{1}(X_i = x) - Y_i P(Y_i|D_i,Z_i,X_i)P(X_i)\Bigg] \\
&= \frac{I(d,z)}{P(D_i = d, Z_i = z|X_i = x)}(y - \mu_d(z,x)) + \mu_d(z,x) - \psi_{d,z} \\
&= \frac{I(d,z)}{\omega_{d,z}(x)}(y - \mu_d(z,x)) + \mu_d(z,x) - \psi_{d,z}
\end{aligned}$$

It follows that if the ΔCBGATE is pathwise differentiable, the influence function of the ΔCBGATE is given by:

$$\begin{aligned}
\phi(h;\theta^{\Delta C},\eta) = &\frac{I(l,v)(y - \mu_l(v,x))}{\omega_{l,v}(x)} + \mu_l(v,x) - \frac{I(m,v)(y - \mu_m(v,x))}{\omega_{m,v}(x)} - \mu_m(v,x) \\
&- \frac{I(l,u)(y - \mu_l(u,x))}{\omega_{l,u}(x)} - \mu_l(u,x) + \frac{I(m,u)(y - \mu_m(u,x))}{\omega_{m,u}(x)} + \mu_m(u,x) - \psi_\theta
\end{aligned}$$

with $\psi_\theta = \mathrm{E}[\mathrm{E}[Y_i|D_i = 1, Z_i = 1, X_i] - \mathrm{E}[Y_i|D_i = 0, Z_i = 1, X_i] - \mathrm{E}[Y_i|D_i = 1, Z_i = 0, X_i] + \mathrm{E}[Y_i|D_i = 0, Z_i = 0, X_i]]$.

### A.3.5 Asymptotic Properties for Estimation with one Propensity Score

The asymptotic properties of the ΔCBGATE estimator with two propensity scores are investigated in this subsection. The following assumptions are imposed:

**Assumption 14.** (Overlap)
*The propensity scores $\lambda_z(w)$ and $\pi_d(z,x)$ are bounded away from 0 and 1:*

$$\kappa < \lambda_z(x), \pi_d(z,x), \hat{\lambda}_z(x), \hat{\pi}_d(z,x) < 1-\kappa \quad \forall x \in \mathcal{X}, z \in \mathcal{Z},$$

*for some $\kappa > 0$.*

**Assumption 15.** (Consistency)
*The estimators of the nuisance functions are sup-norm consistent:*

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z}} |\hat{\mu}_d(z,x) - \mu_d(z,x)| \xrightarrow{p} 0$$
$$\sup_{x\in\mathcal{X},z\in\mathcal{Z}} |\hat{\pi}_d(z,x) - \pi_d(z,x)| \xrightarrow{p} 0$$
$$\sup_{x\in\mathcal{X}} |\hat{\lambda}_z(x) - \lambda_z(x)| \xrightarrow{p} 0$$

**Assumption 16.** (Risk decay)
*The products of the estimation errors for the outcome and propensity models decays as*

$$\mathrm{E}\left[(\hat{\mu}_d(Z_i,X_i) - \mu_d(Z_i,X_i))^2\right]\mathrm{E}\left[(\hat{\pi}_d(Z_i,X_i) - \pi_d(Z_i,X_i))^2\right] = o_p\left(\frac{1}{N}\right)$$
$$\mathrm{E}\left[(\hat{\mu}_d(Z_i,X_i) - \mu_d(Z_i,X_i))^2\right]\mathrm{E}\left[(\hat{\lambda}_z(X_i) - \lambda_z(X_i))^2\right] = o_p\left(\frac{1}{N}\right)$$

*If both nuisance parameters are estimated with the parametric ($\sqrt{N}$-consistent) rate, then the product of the errors would be bounded by $O_p\left(\frac{1}{N^2}\right)$. Hence, it is sufficient for the estimators of the nuisance parameters to be $N^{1/4}$-consistent.*

**Assumption 17.** (Boundness of conditional variances)
*The conditional variances of the outcome is bounded:*

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z}} \mathrm{Var}(Y_i|D_i = d, Z_i = z, X_i = x) < \epsilon_d < \infty$$

The assumptions made are standard in the DML literature (Chernozhukov et al., 2018). Given these assumptions, the following Theorem can be derived:

**Theorem 3.** *Under Assumptions 14 to 17, the proposed estimation strategy for the* $\Delta BGATE$ *obeys*

$$\sqrt{N}(\hat{\theta}^{\Delta C}_{l,m,u,v} - \theta^{\Delta C}_{l,m,u,v}) \xrightarrow{d} N(0, V^\star)$$

*with* $V^\star = \mathrm{E}[\phi^{\Delta C}(H_i; \theta^{\Delta C}, \hat{\eta})^2]$

It follows from Theorem 3 that the estimator is $\sqrt{N}$-consistent and asymptotically normal. The proof looks as follows:

In a first step, define the following terms for easier readability:

$$\begin{aligned}
\pi_d(z,x) &= P(D_i = d | Z_i = z, X_i = x) \\
\mu_d(z,x) &= \mathrm{E}[Y_i | D_i = d, Z_i = z, X_i = x] \\
\lambda_z(x) &= P(Z_i = z | X_i = x) \\
\zeta_{d,z} &= \mathrm{E}\left[\mu_d(z,x) + \frac{I(d,z)(y - \mu_d(z,x))}{\pi_d(z,x)\lambda_z(x)}\right] \\
\Gamma_{d,z}(h) &= \mu_d(z,x) + \frac{I(d,z)(y - \mu_d(z,x))}{\pi_d(z,x)\lambda_z(x)} \\
\theta^{\Delta C}_{l,m,u,v} &= \mathrm{E}[\Gamma_{l,u}(H_i) - \Gamma_{m,u}(H_i) - \Gamma_{l,v}(H_i) + \Gamma_{m,v}(H_i)]
\end{aligned}$$

Due to the linearity in expectations it is enough to focus on $\zeta_{d,z}$, hence we want to show that

$$\begin{aligned}
\sqrt{N}(\hat{\zeta}_{d,z} - \zeta^\star_{d,z}) &\xrightarrow{p} N(0, V^\star) \\
V^\star &= \mathrm{E}\left[\left(\hat{\Gamma}_{d,z}(H_i) - \zeta^\star_{d,z}\right)^2\right]
\end{aligned}$$

with $\zeta^\star_{d,z}$ being an oracle estimator of $\zeta_{d,z}$ if all nuisance functions would be known. Then $\zeta^\star_{d,z}$ is an i.i.d. average, hence:

$$\begin{aligned}
\sqrt{N}(\zeta^\star_{d,z} - \zeta_{d,z}) &\xrightarrow{p} N(0, V^\star) \\
V^\star &= \mathrm{E}\left[\left(\hat{\Gamma}_{d,z}(H_i) - \zeta^\star_{d,z}\right)^2\right]
\end{aligned}$$

**Theorem 4.** *Let* $\mathcal{I}_1$ *and* $\mathcal{I}_2$ *be two half samples such that* $|\mathcal{I}_1| = |\mathcal{I}_1| = \frac{N}{2}$. *Define the estimator* *as follows:*

$$\hat{\zeta}_{d,z} = \frac{|\mathcal{I}_1|}{N}\hat{\zeta}^{\mathcal{I}_1}_{d,z} + \frac{|\mathcal{I}_2|}{N}\hat{\zeta}^{\mathcal{I}_2}_{d,z}$$

$$\hat{\zeta}^{\mathcal{I}_1}_{d,z} = \frac{1}{|\mathcal{I}_1|}\sum_{\mathcal{I}_1}\hat{\Gamma}^{\mathcal{I}_2}_{d,z}(H_i)$$

$$\hat{\zeta}^{\mathcal{I}_2}_{d,z} = \frac{1}{|\mathcal{I}_2|}\sum_{\mathcal{I}_2}\hat{\Gamma}^{\mathcal{I}_1}_{d,z}(H_i)$$

*Then, if Assumptions 5 to 8 it hold, it follows that*

$$\sqrt{N}(\hat{\zeta}_{d,z} - \zeta_{d,z}) \xrightarrow{d} N(0, V^{\star})$$

*with*

$$V^{\star} = \mathrm{E}\left[\left(\hat{\Gamma}_{d,z}(H_i) - \zeta^{\star}_{d,z}\right)^2\right]$$

*Proof.*

$$\begin{aligned}
&\sqrt{N}(\hat{\zeta}_{d,z} - \zeta_{d,z}) \\
&\quad = \sqrt{N}(\hat{\zeta}_{d,z} - \zeta^{\star}_{d,z} + \zeta^{\star}_{d,z} - \zeta_{d,z}) \\
&\quad = \sqrt{N}(\hat{\zeta}_{d,z} - \zeta^{\star}_{d,z}) + \sqrt{N}(\zeta^{\star}_{d,z} - \zeta_{d,z}) \\
&\quad = \sqrt{N}\left(\frac{|\mathcal{I}_1|}{N}\hat{\zeta}^{\mathcal{I}_1}_{d,z} + \frac{|\mathcal{I}_2|}{N}\hat{\zeta}^{\mathcal{I}_2}_{d,z} - \zeta^{\star}_{d,z}\right) + \sqrt{N}(\zeta^{\star}_{d,z} - \zeta_{d,z}) \\
&\quad = \sqrt{N}\left(\frac{|\mathcal{I}_1|}{N}\hat{\zeta}^{\mathcal{I}_1}_{d,z} + \frac{|\mathcal{I}_2|}{N}\hat{\zeta}^{\mathcal{I}_2}_{d,z} - \frac{|\mathcal{I}_1|}{N}\zeta^{\star,\mathcal{I}_1}_{d,z} - \frac{|\mathcal{I}_2|}{N}\zeta^{\star,\mathcal{I}_2}_{d,z}\right) + \sqrt{N}(\zeta^{\star}_{d,z} - \zeta_{d,z}) \\
&\quad = \sqrt{N}\left(\frac{|\mathcal{I}_1|}{N}\hat{\zeta}^{\mathcal{I}_1}_{d,z} - \frac{|\mathcal{I}_1|}{N}\zeta^{\star,\mathcal{I}_1}_{d,z}\right) + \sqrt{N}\left(\frac{|\mathcal{I}_2|}{N}\hat{\zeta}^{\mathcal{I}_2}_{d,z} - \frac{|\mathcal{I}_2|}{N}\zeta^{\star,\mathcal{I}_2}_{d,z}\right) + \sqrt{N}(\zeta^{\star}_{d,z} - \zeta_{d,z})
\end{aligned}$$

The goal is to show that the first two terms converge to zero in probability. It is enough to show this for the first term only, as the same steps can also be directly applied to the remaining second term.

Hence,

$$
\begin{aligned}
&\hat{\zeta}_{d,z}^{\mathcal{I}_1} - \zeta_{d,z}^{\star,\mathcal{I}_1} \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i) + \frac{I(d,z)(Y_i - \hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2})}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}} - \frac{I(d,z)(Y_i - \mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right) \\
&= \underbrace{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)\left(1 - \frac{I(d,z)}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right)\right)}_{\text{Part 1}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(I(d,z)(Y_i - \mu_d(Z_i,X_i))\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i) - \hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\right)}_{\text{Part 2}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(I(d,z)(Y_i - \mu_d(Z_i,X_i))\frac{1}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i) - \hat{\lambda}_z(X_i)^{\mathcal{I}_2})\right)}_{\text{Part 3}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(I(d,z)(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i))\left(\frac{1}{\pi_d(Z_i,X_i)\lambda_z(X_i)} - \frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right)\right)}_{\text{Part 4}}
\end{aligned}
$$

The proof is based on Wager (2020). All four terms converge to zero in probability. For Part 1, after conditioning on $\mathcal{I}_2$, the summands used to build the term are mean-zero and independent. Using the squared $L_2$-norm of Part 1:

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)\left(1 - \frac{I(d,z)}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right)\right)\right)^2\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)\left(1 - \frac{I(d,z)}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right)\right)\right)^2\middle|\mathcal{I}_2\right]\right] \\
&= \mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)\left(1 - \frac{I(d,z)}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right)\middle|\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{Var}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)\left(1 - \frac{I(d,z)}{\pi_d(Z_i,X_i)\lambda_z(X_i)}\right)\middle|\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{E}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)^2\left(\frac{1}{\pi_d(Z_i,X_i)\lambda_z(X_i)} - 1\right)\middle|\mathcal{I}_2\right]\right] \\
&\leqslant \frac{1}{\kappa^2|\mathcal{I}_1|}\mathrm{E}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2} - \mu_d(Z_i,X_i)\right)^2\right] = \frac{o_p(1)}{N}
\end{aligned}
$$

The second equality follows because the summands are mean-zero and independent. The last line follows from Assumption 14 and 15 and the fact that $|\mathcal{I}_1| = N/2$. Hence, Part 1 is $o_p(1/\sqrt{N})$.

Similarly, using the squared $L_2$-norm of Part 2:

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\right)\right)^2\right]\\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\right)\right)^2\middle|I_2\right]\right]\\
&=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\right)\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{Var}\left[\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{E}\left[\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\pi_d(Z_i,X_i)}(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})\right)^2\middle|\mathcal{I}_2\right]\right]\\
&\leqslant\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\mathrm{E}\left[\mathrm{E}\left[(Y_i-\mu_d(Z_i,X_i))^2(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})^2\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\mathrm{E}\left[(Y_i-\mu_d(Z_i,X_i))^2(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})^2\right]\\
&\leqslant\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\epsilon_d\mathrm{E}\left[(\pi_d(Z_i,X_i)-\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2})^2\right]=\frac{o_p(1)}{N}
\end{aligned}
$$

Again, the second equality follows because the summands are mean-zero and independent. The last two inequalities follow from Assumption 15 and 17, the fact that the MSE for the inverse weights decays at the same rate as the MSE for the propensities and the fact that $|\mathcal{I}_1| = N/2$. Hence, Part 2 is $o_p(1/\sqrt{N})$.

Similarly, using the squared $L_2$-norm of Part 3:

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})\right)\right)^2\right] \\
&=\mathrm{E}\left[\mathrm{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})\right)\right)^2\middle| I_2\right]\right] \\
&=\mathrm{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})\right)\middle|\mathcal{I}_2\right]\right] \\
&=\frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{Var}\left[\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})\middle|\mathcal{I}_2\right]\right] \\
&=\frac{1}{|\mathcal{I}_1|}\mathrm{E}\left[\mathrm{E}\left[\left(\frac{I(d,z)(Y_i-\mu_d(Z_i,X_i))}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}\lambda_z(X_i)}(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})\right)^2\middle|\mathcal{I}_2\right]\right] \\
&\leqslant\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\mathrm{E}\left[\mathrm{E}\left[(Y_i-\mu_d(Z_i,X_i))^2(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})^2\middle|\mathcal{I}_2\right]\right] \\
&=\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\mathrm{E}\left[(Y_i-\mu_d(Z_i,X_i))^2(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})^2\right] \\
&\leqslant\frac{1}{\kappa^3|\mathcal{I}_1|}(1-\kappa)^2\epsilon_d\mathrm{E}\left[(\lambda_z(X_i)-\hat{\lambda}_z(X_i)^{\mathcal{I}_2})^2\right]=\frac{o_p(1)}{N}
\end{aligned}
$$

Again, the second equality follows because the summands are mean-zero and independent. The last two inequalities follow from Assumption 15 and 17, the fact that the MSE for the inverse weights decays at the same rate as the MSE for the propensities and the fact that $|\mathcal{I}_1| = N/2$. Hence, Part 3 is $o_p(1/\sqrt{N})$.

Last, using the $L_1$-norm of Part 4:

$$
\begin{aligned}
&\mathrm{E}\left[\left|\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(I(d,z)(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i))\left(\frac{1}{\pi_d(Z_i,X_i)\lambda_z(X_i)}-\frac{1}{\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right)\right)\right|\right]\\
=&\mathrm{E}\left[\left|\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\frac{I(d,z)}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i))\right.\right.\right.\\
&\left.\left.\left.\left(\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)+\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right)\right)\right|\right]\\
\leqslant&\mathrm{E}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left|\frac{I(d,z)}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right|\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\right.\\
&\left.\left|\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)+\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right|\right]\\
=&\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\mathrm{E}\left[\left|\frac{I(d,z)}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right|\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\right.\\
&\left.\left|\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)+\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right|\right]\\
\leqslant&\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\mathrm{E}\left[\left|\frac{I(d,z)}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right|\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\right.\\
&\left.\left(\left|\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)\right|+\left|\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right|\right)\right]\\
=&\mathrm{E}\left[\left|\frac{I(d,z)}{\pi_d(Z_i,X_i)\hat{\lambda}_z(X_i)^{\mathcal{I}_2}}\right|\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\right.\\
&\left.\left(\left|\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)\right|+\left|\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right|\right)\right]\\
\leqslant&\frac{(1-\kappa)^2}{\kappa^2}\mathrm{E}\left[\left|\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right|\left(\left|\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)\right|+\left|\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right|\right)\right]\\
\leqslant&\mathrm{E}\left[\left(\hat{\mu}_d(Z_i,X_i)^{\mathcal{I}_2}-\mu_d(Z_i,X_i)\right)^2\right]^{1/2}\\
&\left(\mathrm{E}\left[\left(\hat{\lambda}_z(X_i)^{\mathcal{I}_2}-\lambda_z(X_i)\right)^2\right]^{1/2}+\mathrm{E}\left[\left(\hat{\pi}_d(Z_i,X_i)^{\mathcal{I}_2}-\pi_d(Z_i,X_i)\right)^2\right]^{1/2}\right)\\
=&\frac{o_p(1)}{\sqrt{N}}
\end{aligned}
$$

The first inequality follows from Cauchy-Schwarz, the fourth line from the triangle inequality, the sixth line from Assumption 14 and the last equality from Assumption 15 and 16 and the fact that $|\mathcal{I}_1| = N/2$. Hence, Part 4 is $o_p(1/\sqrt{N})$.

Hence, we have shown that:

$$\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\hat{\zeta}^{\mathcal{I}_1}_{d,z} - \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\zeta^{\star\mathcal{I}_1}_{d,z} = o_p\left(\frac{1}{\sqrt{N}}\right)$$

Putting all the parts together, we can conclude that:

$$\begin{aligned}&\sqrt{N}(\hat{\zeta}_{d,z} - \zeta_{d,z})\\ &= \underbrace{\sqrt{N}\left(\frac{|\mathcal{I}_{11}|}{N}\hat{\zeta}^{\mathcal{I}_1}_{d,z} - \frac{|\mathcal{I}_1|}{N}\zeta^{\star,\mathcal{I}_1}_{d,z}\right)}_{=o_p(1)} + \underbrace{\sqrt{N}\left(\frac{|\mathcal{I}_2|}{N}\hat{\zeta}^{\mathcal{I}_2}_{d,z} - \frac{|\mathcal{I}_2|}{N}\zeta^{\star,\mathcal{I}_2}_{d,z}\right)}_{=o_p(1)} + \underbrace{\sqrt{N}(\zeta^{\star}_{d,z} - \zeta_{d,z})}_{\xrightarrow{d} N(0,V^{\star})}\end{aligned}$$

Hence, the estimator $\hat{\theta}^{\Delta C}_{l,m,u,v}$ is $\sqrt{N}$-consistent and asymptotically normal.

# B Appendix: Estimation Procedures

Algorithm 2 shows how a propensity score can be truncated and normalized for better finite sample properties. This algorithm can be used for all propensity scores in this paper, namely $\lambda_z(w)$, $\lambda_z(x)$, $\pi_d(z,x)$ and $\omega_{d,z}(x)$. From now on, the propensity score is called $\pi_d(z,x)$. Be aware that the score function has to be adapted because weight$_{norm,i}$ replaces $\frac{I(d)}{\pi_d(z,x)}$ and not only $\pi_d(z,x)$.

---

**Algorithm 2:** NORMALIZATION AND TRUNCATION OF PROPENSITY SCORE WEIGHTS

---

**Input :** Data: $\{w_i, z_i, y_i\}$
Propensity score: $\pi_d(z_i, x_i)$

**Output:** Normalized and truncated weights: weight$_{norm,i}$

**begin**
- Number of observations: $N$
- **for** $i$ = 1, ..., $N$ **do**
  - TRUNCATION:
  - $\pi_d(z_i, x_i) = \max(\pi_d(z_i, x_i), 0.0001)$
  - DEFINITION WEIGHTS:
  - $weight = \frac{I(d)}{\pi_d(z_i, x_i)}$
- **end**
- **Define:** Sum of the weights: $weight_{sum} = \sum_{i=1}^{N} weight_i$
- **for** $i$ = 1, ..., $N$ **do**
  - NORMALIZATION:
  - $weight_{norm,i} = \frac{weight_i}{weight_{sum}}$
  - TRUNCATION:
  - $weight_{norm,i} = \min(weight_{norm,i}, 0.05)$
- **end**
- **Define:** Sum of the new weights: $weight_{norm,sum} = \sum_{i=1}^{N} weight_{norm,i}$
- **for** $i$ = 1, ..., $N$ **do**
  - NORMALIZATION:
  - $weight_{norm,i} = \frac{weight_{norm,i}}{weight_{norm,sum}} \cdot N$
- **end**

**end**

---

Algorithm 3 shows the algorithm to estimate the ΔBGATE with discrete treatment and moderator variables.

**Algorithm 3:** DML FOR ΔBGATE

**Input** : Data: $h_i = \{x_i, z_i, d_i, y_i\}$

**Output:** ΔBGATE $= \hat{\theta}^{\Delta B}_{l,m,u,v}$

**begin**

**create folds:** Split sample into K random folds $(S_k)_{k=1}^{K}$ of observations $\{1, \ldots, \frac{N}{K}\}$ with size of each fold $\frac{N}{K}$. Define $S_k^c := \{1, \ldots, N\} \backslash \{S_k\}$

**for** $k$ in $\{1, \ldots, K\}$ **do**

**for** $d$ in $\{l, m\}$ **do**

RESPONSE FUNCTIONS:

estimate: $\hat{\mu}_d(z_i, x_i) = \hat{\mathrm{E}}[Y_i | D_i = d_i, X_i = x_i, Z_i = z_i]$ in $\{x_i, y_i, z_i\}_{i \in S_k^c, d_i = d}$

PROPENSITY SCORE:

estimate: $\hat{\pi}_d(x_i, z_i) = \hat{P}(D_i = d_i | X_i = x_i, Z_i = z_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

**end**

PSEUDO-OUTCOME:

estimate: $\hat{\delta}_{l,m}(h_i) = \hat{\mu}_l(z_i, x_i) - \hat{\mu}_m(z_i, x_i) + \frac{I(l)(y_i - \hat{\mu}_l(z_i, x_i))}{\hat{\pi}_l(x_i, z_i)} - \frac{I(m)(y_i - \hat{\mu}_m(z_i, x_i))}{\hat{\pi}_m(x_i, z_i)}$ in $\{h_i\}_{i \in S_k}$

**end**

**create folds:** Split sample $S_k$ into J random folds $(S_j)_{j=1}^{J}$ of observations $\{1, \ldots, \frac{N}{K \cdot J}\}$ with size of each fold $\frac{N}{K \cdot J}$. Define $S_j^c := \{1, \ldots, \frac{N}{K}\} \backslash \{S_j\}$

**for** $j$ in $\{1, \ldots, J\}$ **do**

**for** $z$ in $\{v, u\}$ **do**

PSEUDO-OUTCOME REGRESSION:

estimate: $\hat{g}_z(w_i) = \hat{\mathrm{E}}[\hat{\delta}_{l,m}(h_i) | Z_i = z_i, W_i = w_i]$ in $\{h_i\}_{i \in S_j^c, z_i = z}$

PROPENSITY SCORE:

estimate: $\hat{\lambda}_z(w_i) = \hat{P}(Z_i = z_i | W_i = w_i)$ in $\{w_i, z_i\}_{i \in S_j^c}$

**end**

ΔBGATE FUNCTION:

EFFECT:

estimate:

$\hat{\theta}^{\Delta B}_{j,l,m,u,v} = \frac{K \cdot J}{N} \sum_{i \in S_j} \left[ \hat{g}_u(W_i) - \hat{g}_v(W_i) + \frac{I(u)(\hat{\delta}_{l,m}(H_i) - \hat{g}_u(W_i))}{\hat{\lambda}_u(W_i)} - \frac{I(v)(\hat{\delta}_{l,m}(H_i) - \hat{g}_v(W_i))}{\hat{\lambda}_v(W_i)} \right]$

STANDARD ERRORS:

estimate:

$\hat{\theta}^{\Delta B_{SE}}_{l,m,u,v,j} = \frac{K \cdot J}{N} \sum_{i \in S_j} \left[ \left( \hat{g}_u(w_i) - \hat{g}_v(w_i) + \frac{I(u)(\delta_{l,m}(h_i) - \hat{g}_u(w_i))}{\hat{\lambda}_u(w_i)} - \frac{I(v)(\hat{\delta}_{l,m}(h_i) - \hat{g}_v(w_i))}{\hat{\lambda}_v(W_i)} \right)^2 \right]$

**end**

estimate effect: $\hat{\theta}^{\Delta B}_{l,m,u,v} = \frac{1}{J \cdot K} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\theta}^{\Delta B}_{j,k,l,m,u,v}$

estimate standard errors: $SE(\hat{\theta}^{\Delta B}_{l,m,u,v}) = \sqrt{\frac{1}{J \cdot K} \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\theta}^{\Delta B_{SE}}_{j,k,l,m,u,v} - \left( \hat{\theta}^{\Delta B}_{j,k,l,m,u,v} \right)^2}$

**end**

Algorithm 4 shows the suggested estimation procedure for the ΔCBGATE. It is similar to the estimation of an ATE but with a different score function.

**Algorithm 4:** DML FOR ΔCBGATE

**Input :** Data: $h_i = \{x_i, z_i, d_i, y_i\}$

**Output:** ΔCBGATE = $\hat{\theta}^{\Delta C}$

**begin**

**create folds:** Split sample into K random folds $(S_k)_{k=1}^K$ of observations $\{1, \ldots, \frac{N}{K}\}$ with size of each fold $\frac{N}{K}$. Also define $S_k^c := \{1, \ldots, N\} \backslash S_k$

**for** $k$ in $\{1, \ldots, K\}$ **do**

**for** $d$ in $\{l, m\}$ **do**

**for** $z$ in $\{v, u\}$ **do**

RESPONSE FUNCTIONS:

estimate: $\hat{\mu}_d(z_i, x_i) = \hat{\mathrm{E}}[Y_i | X_i = x_i, D_i = d_i, Z_i = z_i]$ in $\{x_i, y_i, z_i\}_{i \in S_k^c, d_i = d, z_i = z}$

PROPENSITY SCORE:

**VERSION 1:**

estimate: $\hat{\omega}_{d,z}(x_i) = \hat{P}(D_i = d_i, Z_i = z_i | X_i = x_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

**VERSION 2:**

estimate: $\hat{\pi}_d(x_i, z_i) = \hat{P}(D_i = d_i | X_i = x_i, Z_i = z_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

estimate: $\hat{\lambda}_z(x_i) = \hat{P}(Z_i = z_i | X_i = x_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

calculate: $\hat{\omega}_{d,z}(x_i) = \hat{\pi}_d(x_i, z_i) \cdot \hat{\lambda}_z(x_i)$

**end**

**end**

ΔCBGATE FUNCTION:

EFFECT:

estimate: $\hat{\theta}_{k,l,m,u,v}^{\Delta C} = \frac{K}{N} \sum_{i \in S_k} \Big[ \hat{\mu}_l(v, x_i) - \hat{\mu}_m(v, x_i) - \hat{\mu}_l(u, x_i) + \hat{\mu}_m(u, x_i) + \frac{I(l,v)(y_i - \hat{\mu}_l(v,x_i))}{\hat{\omega}_{l,v}(x_i)} - \frac{I(m,v)(y_i - \hat{\mu}_m(v,x_i))}{\hat{\omega}_{m,v}(x_i)} - \frac{I(l,u)(y_i - \hat{\mu}_l(u,x_i))}{\hat{\omega}_{l,u}(x_i)} + \frac{I(m,u)(y_i - \hat{\mu}_m(u,x_i))}{\hat{\omega}_{m,u}(x_i)} \Big]$

STANDARD ERRORS:

estimate: $\hat{\theta}_{k,l,m,u,v}^{\Delta C,SE} = \frac{K}{N} \sum_{i \in S_k} \Big[ \Big( \hat{\mu}_l(v, x_i) - \hat{\mu}_m(v, x_i) - \hat{\mu}_l(u, x_i) + \hat{\mu}_m(u, x_i) + \frac{I(l,v)(y_i - \hat{\mu}_l(v,x_i))}{\hat{\omega}_{l,v}(x_i)} - \frac{I(m,v)(y_i - \hat{\mu}_m(v,x_i))}{\hat{\omega}_{m,v}(x_i)} - \frac{I(l,u)(y_i - \hat{\mu}_l(u,x_i))}{\hat{\omega}_{l,u}(x_i)} + \frac{I(m,u)(y_i - \hat{\mu}_m(u,x_i))}{\hat{\omega}_{m,u}(x_i)} \Big)^2 \Big]$

**end**

estimate effect: $\hat{\theta}_{l,m,u,v}^{\Delta C} = \frac{1}{K} \sum_{k=1}^K \hat{\theta}_{k,l,m,u,v}^{\Delta C}$

estimate standard errors: $SE(\hat{\theta}^{\Delta C}) = \sqrt{\frac{1}{K} \sum_{k=1}^K \hat{\theta}_{k,l,m,u,v}^{\Delta C,SE} - \left( \hat{\theta}_{l,m,u,v}^{\Delta C} \right)^2}$

**end**

Algorithm 5 shows the procedure for estimating $\theta^{\Delta G}$ with DML by Chernozhukov et al. (2018).

Summarized, estimate an average treatment effect (ATE) in groups $Z_i = 1$ and $Z_i = 0$ separately. Then take the difference of those two effects to receive $\hat{\theta}^{\Delta G}$.

---

**Algorithm 5:** DML FOR ΔGATE

---

**Input :** Data: $h_i = \{x_i, d_i, z_i, y_i\}$

**Output:** ΔGATE: $\hat{\theta}^{\Delta G}$

**begin**

  **for** $z$ in $\{v, u\}$ **do**

    **create folds:** Split sample into K random folds $(S_k)_{k=1}^{K}$ of observations $\{1, \ldots, \frac{N}{K}\}$ with size of each fold $\frac{N}{K}$. Also define $S_k^c := \{1, \ldots, N\} \backslash S_k$

    **for** $k$ in $\{1, \ldots, K\}$ **do**

      **for** $d$ in $\{l, m\}$ **do**

        RESPONSE FUNCTIONS:

        estimate: $\hat{\mu}_d(z_i, x_i) = \hat{\mathrm{E}}[Y_i | D_i = d_i, X_i = x_i, Z_i = z_i]$ in $\{x_i, z_i, y_i\}_{i \in S_k^c, d_i = d}$

        PROPENSITY SCORE:

        estimate: $\hat{\pi}_d(z_i, x_i) = \hat{P}(D_i = d_i | X_i = x_i, Z_i = z_i)$ in $\{x_i, d_i, z_i\}_{i \in S_k^c}$

      **end**

      ATE FUNCTION:

      EFFECT:

      estimate: $\hat{\theta}_{k,l,m} =$
      $\frac{K}{N} \sum_{i \in S_k} \left[ \hat{\mu}_l(z_i, x_i) - \hat{\mu}_m(z_i, x_i) + \frac{\mathcal{I}(D_i = l)(Y_i - \hat{\mu}_l(z_i, x_i))}{\hat{\pi}_l(z_i, x_i)} - \frac{\mathcal{I}(D_i = m)(Y_i - \hat{\mu}_m(z_i, x_i))}{\hat{\pi}_m(z_i, x_i)} \right]$

      STANDARD ERRORS:

      estimate: $\hat{\theta}_{k,l,m}^{SE} =$
      $\frac{K}{N} \sum_{i \in S_k} \left[ \left( \hat{\mu}_l(z_i, x_i) - \hat{\mu}_m(z_i, x_i) + \frac{\mathcal{I}(d_i = l)(y_i - \hat{\mu}_l(z_i, x_i))}{\hat{\pi}_l(z_i, x_i)} - \frac{\mathcal{I}(d_i = m)(y_i - \hat{\mu}_m(z_i, x_i))}{(\hat{\pi}_m(z_i, x_i))} \right)^2 \right]$

    **end**

    estimate effect: $\hat{\theta}_{l,m}(z) = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{k,l,m}$

    estimate standard errors: $\mathrm{Var}(\hat{\theta}_{l,m}(z)) = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{k,l,m}^{SE} - \hat{\theta}_{l,m}(z)^2$

  **end**

  calculate effect: $\hat{\theta}_{l,m,u,v}^{\Delta G} = \hat{\theta}_{l,m}(u) - \hat{\theta}_{l,m}(v)$

  calculate standard errors: $SE(\hat{\theta}_{l,m,u,v}^{\Delta G}) = \sqrt{\mathrm{Var}(\hat{\theta}_{l,m}(u)) + \mathrm{Var}(\hat{\theta}_{l,m}(v))}$

**end**

---

# C Appendix: Monte Carlo Study

This section explains details of the Monte Carlo study conducted for the ΔBGATE and the ΔCBGATE. For simplicity, the simulations cover the case where $Z_i$ and $D_i$ are binary variables.

## C.1 Performance Measures

For the evaluation of the different estimators, different measures are considered. First, we look at the measures to evaluate the estimation of the effects. The performance measures are shown for $\hat{\theta}^{\Delta B}_{l,m,u,v}$ only, but the same performance measures are also used for $\hat{\theta}^{\Delta C}_{l,m,u,v}$ and $\hat{\theta}^{\Delta G}_{l,m,u,v}$.

$$
\begin{aligned}
BIAS(\hat{\theta}^{\Delta B}_{l,m,u,v}) &= \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{\Delta B}_{r,l,m,u,v} - \theta^{\Delta B}_{l,m,u,v} \\
|BIAS(\hat{\theta}^{\Delta B}_{l,m,u,v})| &= \frac{1}{R}\sum_{r=1}^{R}|\hat{\theta}^{\Delta B}_{r,l,m,u,v} - \theta^{\Delta B}_{l,m,u,v}| \\
SD(\hat{\theta}^{\Delta B}_{l,m,u,v}) &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{\Delta B}_{r,l,m,u,v} - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{\Delta B}_{r,l,m,u,v})^2} \\
RMSE(\hat{\theta}^{\Delta B}_{l,m,u,v}) &= \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{\Delta B}_{r,l,m,u,v} - \theta^{\Delta B}_{l,m,u,v})^2} \\
SKEWNESS(\hat{\theta}^{\Delta B}_{l,m,u,v}) &= \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\hat{\theta}^{\Delta B}_{l,m,u,v} - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{\Delta B}_{r,l,m,u,v}}{SD(\hat{\theta}^{\Delta B}_{l,m,u,v})}\right)^3 \\
Ex.KURTOSIS(\hat{\theta}^{\Delta B}_{l,m,u,v}) &= \frac{1}{R}\sum_{i=1}^{R}\left(\frac{\hat{\theta}^{\Delta B}_{r,l,m,u,v} - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{\Delta B}_{r,l,m,u,v}}{SD(\hat{\theta}^{\Delta B}_{l,m,u,v})}\right)^4 - 3
\end{aligned}
$$

Furthermore, the following performance measures are used to evaluate the inference procedures:

$$
\begin{aligned}
BIAS(\widehat{SE}(\hat{\theta}^{\Delta B}_{l,m,u,v})) &= \frac{1}{R}\sum_{i=1}^{R}SE^{r}(\hat{\theta}^{\Delta B}_{r,l,m,u,v}) - SD(\hat{\theta}^{\Delta B}_{l,m,u,v}) \\
CovP(\alpha) &= P\Bigg[\left(\theta^{\Delta B}_{l,m,u,v} > \frac{1}{R}\sum_{i=1}^{r}(\hat{\theta}^{\Delta B}_{r,l,m,u,v} - Z_{\frac{\alpha}{2}}SE(\hat{\theta}^{\Delta B}_{r,l,m,u,v}))\right) \\
&\quad \cup \left(\theta_{l,m,u,v} < \frac{1}{R}\sum_{i=1}^{r}(\hat{\theta}^{\Delta B}_{r,l,m,u,v} + Z_{1-\frac{\alpha}{2}}SE(\hat{\theta}^{\Delta B}_{r,l,m,u,v}))\right)\Bigg]
\end{aligned}
$$

## C.2 Details for ∆BGATE

### C.2.1 Random Forest Tuning Parameters

The optimal hyperparameters for the different random forests, namely the tree depth and the minimum leaf size, are tuned by a grid search (maximum tree depth: [5, 10, 15, 20], minimum leaf size: [2, 5, 10, 20]) using a random forest with 1000 trees and three folds. They are tuned using five different data draws and then fixed for the whole simulation. The optimal combination of maximal tree depth and minimum leaf size is chosen by taking the combination that appears most often in the five draws. Table C.1 shows the optimal hyperparameters for the different DGPs and sample sizes.

**Table C.1:** Simulation Study: Optimal hyperparameters for random forests

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP: non-linear with linear heterogeneous treatment effects | | | | | | | | | | | | |
| | N = 2,500 | | | | | | N = 10,000 | | | | | |
| | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 5 | 5 | 5 | 10 | 5 | 5 | 10 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 10 | 2 | 2 | 20 | 20 | 20 | 20 | 2 | 20 | 20 | 5 | 20 |
| DGP: non-linear with non-linear heterogeneous treatment effects | | | | | | | | | | | | |
| | N = 2,500 | | | | | | N = 10,000 | | | | | |
| | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 10 | 5 | 5 | 5 | 10 | 5 | 10 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 5 | 2 | 2 | 20 | 20 | 20 | 10 | 2 | 20 | 20 | 2 | 20 |
| DGP: non-linear with non-linear heterogeneous treatment effects and $Z_i$ influencing $X_{i,5}$ | | | | | | | | | | | | |
| | N = 2,500 | | | | | | N = 10,000 | | | | | |
| | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ | $\mu_1$ | $\mu_0$ | $\pi_d$ | $g_1$ | $g_0$ | $\lambda_z$ |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 10 | 5 | 5 | 5 | 10 | 5 | 10 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 10 | 2 | 20 | 20 | 20 | 20 | 10 | 2 | 20 | 20 | 20 | 20 |

*Note:* This table depicts the optimal hyperparameters for the random forests. The grid values are as follows: maximum tree depth: [5, 10, 15, 20], minimum leaf size: [2, 5, 10, 20].

### C.2.2 Results

Table C.2 and Table C.3 show the results for a non-linear DGP with linear heterogeneity with $N = 2,500$ and $N = 10,000$, respectively. The first part of the tables depicts the results of estimating the outcome regression in the second estimation step with a linear regression. The second part of the tables shows the results of using a random forest to estimate the outcome regressions in the second step. Using a linear regression works well when the heterogeneities are linear. However, using a random forest works equally well. Furthermore, the tables show that the standard error halves by increasing the sample size by four. This suggests a $\sqrt{N}$-convergence rate of the estimator also in finite samples.

**Table C.2:** Simulation results: Non-linear DGP with a linear treatment effect (N = 2,500)

| | DGP: non-linear with linear heterogeneous treatment effects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| | | | | **2nd estimation step with linear regression** | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.596 | 0.001 | 0.079 | 0.097 | 0.097 | -0.015 | -0.066 | 0.001 | 95.70 | 81.20 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.493 | -0.005 | 0.091 | 0.115 | 0.115 | -0.062 | 0.069 | 0.002 | 96.00 | 80.90 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.560 | -0.009 | 0.093 | 0.117 | 0.117 | -0.157 | 0.339 | 0.001 | 96.40 | 80.40 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.596 | -0.004 | 0.080 | 0.099 | 0.100 | -0.026 | -0.032 | 0.000 | 95.20 | 80.60 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.597 | -0.004 | 0.079 | 0.100 | 0.100 | 0.069 | 0.271 | -0.000 | 94.90 | 81.90 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.449 | -0.007 | 0.103 | 0.129 | 0.129 | -0.102 | 0.117 | 0.001 | 96.00 | 81.40 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.560 | -0.012 | 0.090 | 0.112 | 0.113 | -0.111 | -0.124 | 0.002 | 95.70 | 79.70 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.450 | -0.007 | 0.100 | 0.127 | 0.127 | -0.090 | 0.207 | 0.001 | 95.50 | 81.20 |
| | | | | **2nd estimation step with random forest** | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.596 | 0.001 | 0.079 | 0.097 | 0.097 | -0.015 | -0.066 | 0.001 | 95.70 | 81.20 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.493 | -0.006 | 0.092 | 0.115 | 0.115 | -0.076 | 0.074 | 0.003 | 96.00 | 81.70 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.560 | -0.008 | 0.093 | 0.117 | 0.117 | -0.114 | 0.361 | 0.001 | 96.00 | 80.50 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.596 | -0.004 | 0.081 | 0.101 | 0.101 | -0.035 | -0.058 | -0.000 | 95.00 | 78.90 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.597 | -0.004 | 0.079 | 0.100 | 0.100 | 0.091 | 0.313 | 0.000 | 95.00 | 81.70 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.449 | -0.006 | 0.104 | 0.130 | 0.131 | -0.096 | 0.132 | 0.001 | 96.00 | 81.30 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.560 | -0.011 | 0.091 | 0.113 | 0.114 | -0.092 | -0.106 | 0.002 | 95.60 | 79.70 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.450 | -0.008 | 0.102 | 0.128 | 0.129 | -0.066 | 0.177 | 0.001 | 95.00 | 80.50 |

*Note:* This table shows results for $N = 2,500$, $R = 1,000$ and a non-linear DGP with linear heterogeneous treatment effects in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

**Table C.3:** Simulation results: Non-linear DGP with a linear treatment effect (N = 10,000)

| | DGP: non-linear with linear heterogeneous treatment effects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| | | | | **2nd estimation step with linear regression** | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.596 | 0.005 | 0.038 | 0.046 | 0.046 | -0.101 | -0.062 | 0.002 | 96.40 | 81.60 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.493 | 0.004 | 0.043 | 0.054 | 0.054 | -0.193 | 0.065 | 0.002 | 97.60 | 80.00 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.560 | 0.001 | 0.043 | 0.053 | 0.053 | 0.073 | -0.455 | 0.003 | 97.20 | 83.20 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.596 | 0.004 | 0.037 | 0.045 | 0.045 | -0.126 | 0.040 | 0.003 | 95.60 | 84.40 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.597 | 0.004 | 0.037 | 0.046 | 0.046 | -0.108 | 0.012 | 0.002 | 96.00 | 83.60 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.449 | 0.008 | 0.049 | 0.060 | 0.060 | -0.274 | 0.013 | 0.003 | 96.40 | 83.20 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.560 | 0.001 | 0.043 | 0.053 | 0.053 | 0.012 | -0.400 | 0.003 | 98.00 | 82.40 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.450 | 0.008 | 0.048 | 0.059 | 0.059 | -0.354 | 0.199 | 0.003 | 95.60 | 81.20 |
| | | | | **2nd estimation step with random forest** | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.596 | 0.005 | 0.038 | 0.046 | 0.046 | -0.101 | -0.062 | 0.002 | 96.40 | 81.60 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.493 | 0.005 | 0.043 | 0.054 | 0.055 | -0.171 | 0.101 | 0.002 | 96.80 | 82.00 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.560 | 0.001 | 0.043 | 0.053 | 0.053 | 0.016 | -0.448 | 0.004 | 98.00 | 83.60 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.596 | 0.004 | 0.037 | 0.045 | 0.045 | -0.153 | 0.004 | 0.003 | 96.40 | 84.40 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.597 | 0.004 | 0.038 | 0.046 | 0.046 | -0.125 | 0.053 | 0.002 | 95.60 | 83.20 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.449 | 0.009 | 0.049 | 0.060 | 0.061 | -0.242 | -0.061 | 0.003 | 95.60 | 82.00 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.560 | 0.001 | 0.043 | 0.053 | 0.053 | -0.001 | -0.354 | 0.003 | 96.80 | 82.80 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.450 | 0.008 | 0.048 | 0.059 | 0.060 | -0.311 | 0.215 | 0.003 | 95.20 | 82.00 |

*Note:* This table shows results for $N = 10,000$, $R = 250$ and a non-linear DGP with linear heterogeneous treatment effects in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

Table C.4 and Table C.5 show the result for a non-linear DGP with non-linear heterogeneous treatment effects. Using a linear regression to estimate the outcome regressions in the second step leads to biased results if we want to balance the distribution of the $X_i$'s that lead to those heterogeneous treatment effects. This is the case because the heterogeneities are non-linear. However, due to the double-robust property of the DML estimator, the bias is not huge. Using DML with a random forest in the second step works well. Again, the standard errors are halved by increasing the sample size by four.

**Table C.4:** Simulation results: Non-linear DGP with a non-linear treatment effect (N $= 2,500$)

| DGP: non-linear with non-linear heterogeneous treatment effects | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| **2nd estimation step with linear regression** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.272 | -0.009 | 0.084 | 0.104 | 0.104 | 0.004 | -0.012 | -0.000 | 95.20 | 79.30 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.378 | -0.012 | 0.097 | 0.122 | 0.122 | -0.122 | 0.260 | 0.004 | 96.80 | 82.30 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.349 | -0.014 | 0.099 | 0.125 | 0.126 | -0.141 | 0.372 | 0.000 | 95.30 | 80.60 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.270 | -0.008 | 0.085 | 0.106 | 0.106 | -0.078 | 0.026 | -0.000 | 95.70 | 79.60 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.272 | -0.010 | 0.085 | 0.107 | 0.107 | 0.009 | 0.273 | -0.001 | 94.00 | 80.30 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.492 | -0.016 | 0.108 | 0.134 | 0.135 | -0.112 | 0.106 | 0.003 | 96.00 | 80.50 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.349 | -0.019 | 0.096 | 0.120 | 0.121 | -0.148 | -0.023 | 0.002 | 95.20 | 79.60 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.503 | -0.024 | 0.107 | 0.133 | 0.135 | -0.111 | 0.220 | 0.003 | 94.40 | 81.40 |
| **2nd estimation step with random forest** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.272 | -0.009 | 0.084 | 0.104 | 0.104 | 0.004 | -0.012 | -0.000 | 95.20 | 79.30 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.378 | -0.004 | 0.096 | 0.122 | 0.122 | -0.122 | 0.192 | 0.002 | 96.00 | 82.00 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.349 | -0.012 | 0.099 | 0.126 | 0.126 | -0.100 | 0.387 | 0.001 | 95.30 | 81.50 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.270 | -0.009 | 0.086 | 0.107 | 0.108 | -0.071 | 0.006 | -0.001 | 95.20 | 79.60 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.272 | -0.010 | 0.085 | 0.107 | 0.108 | 0.028 | 0.309 | -0.001 | 94.30 | 80.60 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.492 | -0.024 | 0.110 | 0.136 | 0.138 | -0.144 | 0.069 | 0.000 | 95.00 | 79.50 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.349 | -0.018 | 0.097 | 0.121 | 0.122 | -0.123 | -0.007 | 0.002 | 94.90 | 80.50 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.503 | -0.033 | 0.109 | 0.134 | 0.138 | -0.113 | 0.138 | 0.000 | 94.30 | 79.20 |

*Note:* This table shows results for $N = 2,500$, $R = 1,000$ and a non-linear DGP with a non-linear treatment effect in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

Table C.6 and Table C.7 show the result for a non-linear DGP with non-linear heterogeneous treatment effects and some covariates that are influenced by the moderator variable $Z_i$. Similarly, as seen in the setting with only non-linear heterogeneous treatment effects, using a linear regression in the second step leads to biased results. Using a random forest in the second step works well, even if the moderator variable influences some confounders.

**Table C.5:** Simulation results: Non-linear DGP with a non-linear treatment effect (N = $10,000$)

| DGP: non-linear with non-linear heterogeneous treatment effects | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| **2nd estimation step with linear regression** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.272 | -0.005 | 0.041 | 0.049 | 0.050 | -0.044 | -0.293 | 0.001 | 95.20 | 80.80 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.378 | -0.017 | 0.048 | 0.058 | 0.060 | -0.251 | -0.202 | 0.003 | 95.20 | 80.40 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.349 | -0.005 | 0.047 | 0.056 | 0.056 | 0.045 | -0.659 | 0.004 | 97.60 | 82.00 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.270 | -0.003 | 0.040 | 0.048 | 0.049 | -0.012 | -0.327 | 0.002 | 95.60 | 83.60 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.272 | -0.006 | 0.041 | 0.049 | 0.050 | -0.046 | -0.347 | 0.001 | 96.80 | 80.80 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.492 | -0.022 | 0.054 | 0.064 | 0.067 | -0.268 | -0.078 | 0.003 | 94.40 | 79.20 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.349 | -0.005 | 0.046 | 0.055 | 0.056 | 0.018 | -0.698 | 0.003 | 98.80 | 79.60 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.503 | -0.030 | 0.055 | 0.063 | 0.070 | -0.291 | -0.066 | 0.003 | 94.40 | 78.00 |
| **2nd estimation step with random forest** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.272 | -0.005 | 0.041 | 0.049 | 0.050 | -0.044 | -0.293 | 0.001 | 95.20 | 80.80 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.378 | 0.001 | 0.045 | 0.057 | 0.057 | -0.174 | -0.258 | 0.002 | 96.80 | 78.80 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.349 | -0.003 | 0.047 | 0.056 | 0.056 | 0.041 | -0.614 | 0.004 | 97.60 | 83.60 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.270 | -0.004 | 0.040 | 0.048 | 0.049 | -0.033 | -0.337 | 0.002 | 96.40 | 82.80 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.272 | -0.005 | 0.041 | 0.050 | 0.050 | -0.057 | -0.328 | 0.001 | 96.00 | 79.60 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.492 | -0.016 | 0.052 | 0.063 | 0.065 | -0.233 | -0.018 | 0.002 | 94.80 | 79.60 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.349 | -0.004 | 0.046 | 0.055 | 0.055 | 0.002 | -0.651 | 0.004 | 98.40 | 79.60 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.503 | -0.026 | 0.054 | 0.063 | 0.068 | -0.297 | 0.043 | 0.002 | 94.80 | 78.40 |

*Note:* This table shows results for $N = 10,000$, $R = 250$ and a non-linear DGP with a non-linear treatment effect in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

**Table C.6:** Simulation results: Non-linear DGP with a non-linear treatment effect and a moderator influencing some covariates (N = $2,500$)

| DGP: non-linear with non-linear treatment effect and $Z_i$ influencing $X_{i,5}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| **2nd estimation step with linear regression** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.425 | -0.004 | 0.083 | 0.104 | 0.105 | -0.058 | 0.182 | 0.002 | 95.10 | 80.70 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.532 | -0.016 | 0.100 | 0.126 | 0.127 | -0.031 | 0.200 | 0.003 | 95.50 | 80.80 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.501 | -0.021 | 0.102 | 0.127 | 0.129 | -0.002 | 0.089 | -0.001 | 94.10 | 78.60 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.423 | -0.013 | 0.086 | 0.108 | 0.108 | -0.014 | 0.093 | -0.000 | 93.90 | 80.60 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.288 | -0.019 | 0.099 | 0.123 | 0.124 | 0.003 | -0.029 | -0.001 | 94.40 | 77.60 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.645 | -0.019 | 0.109 | 0.136 | 0.137 | -0.065 | -0.146 | 0.004 | 94.50 | 79.90 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.502 | -0.024 | 0.101 | 0.125 | 0.127 | -0.080 | -0.001 | -0.001 | 94.10 | 79.40 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.655 | -0.027 | 0.110 | 0.135 | 0.138 | -0.109 | -0.198 | 0.003 | 94.60 | 79.80 |
| **2nd estimation step with random forest** | | | | | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.425 | -0.004 | 0.083 | 0.104 | 0.105 | -0.058 | 0.182 | 0.002 | 95.10 | 80.70 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.532 | -0.007 | 0.099 | 0.125 | 0.125 | 0.005 | 0.126 | 0.001 | 95.50 | 79.90 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.501 | -0.020 | 0.102 | 0.127 | 0.129 | -0.001 | 0.076 | -0.000 | 94.70 | 78.50 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.423 | -0.014 | 0.086 | 0.108 | 0.109 | -0.028 | 0.068 | 0.000 | 94.10 | 80.20 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.288 | -0.020 | 0.099 | 0.123 | 0.125 | 0.004 | -0.030 | -0.001 | 94.20 | 77.10 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.645 | -0.026 | 0.110 | 0.137 | 0.139 | -0.072 | -0.148 | 0.002 | 93.70 | 79.60 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.502 | -0.024 | 0.100 | 0.124 | 0.126 | -0.081 | 0.022 | -0.000 | 94.20 | 79.50 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.655 | -0.036 | 0.111 | 0.135 | 0.140 | -0.115 | -0.170 | 0.001 | 93.90 | 79.50 |

*Note:* This table shows results for $N = 2,500$, $R = 1,000$ for a non-linear DGP with non-linear heterogeneous treatment effects and a covariate influenced by the moderator variable in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

**Table C.7:** Simulation results: Non-linear DGP with a non-linear treatment effect and a moderator influencing some covariates (N = $10,000$)

| | DGP: non-linear with non-linear treatment effect and $Z_i$ influencing $X_{i,5}$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Estimation of effects | | | | | | | Estimation of standard errors | | |
| | 2nd step | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| | | | | **2nd estimation step with linear regression** | | | | | | | |
| $\theta^{\Delta G}$ | DML with OLS | 0.425 | -0.005 | 0.040 | 0.050 | 0.050 | -0.074 | -0.300 | 0.002 | 96.80 | 81.20 |
| $\theta^{\Delta B}(X_0)$ | DML with OLS | 0.532 | -0.024 | 0.050 | 0.058 | 0.063 | -0.102 | -0.270 | 0.004 | 94.80 | 77.60 |
| $\theta^{\Delta B}(X_1)$ | DML with OLS | 0.501 | -0.012 | 0.048 | 0.060 | 0.061 | -0.027 | 0.139 | 0.001 | 94.80 | 79.20 |
| $\theta^{\Delta B}(X_2)$ | DML with OLS | 0.423 | -0.008 | 0.040 | 0.050 | 0.051 | 0.001 | -0.088 | 0.002 | 94.80 | 79.20 |
| $\theta^{\Delta B}(X_5)$ | DML with OLS | 0.288 | -0.012 | 0.050 | 0.061 | 0.062 | 0.103 | -0.129 | -0.001 | 94.80 | 78.00 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with OLS | 0.645 | -0.028 | 0.055 | 0.063 | 0.069 | 0.000 | 0.047 | 0.005 | 93.60 | 81.20 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with OLS | 0.502 | -0.012 | 0.048 | 0.060 | 0.061 | -0.027 | 0.181 | 0.000 | 94.80 | 79.20 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with OLS | 0.655 | -0.035 | 0.058 | 0.063 | 0.072 | -0.010 | 0.096 | 0.004 | 92.80 | 77.60 |
| | | | | **2nd estimation step with random forest** | | | | | | | |
| $\theta^{\Delta G}$ | DML with RF | 0.425 | -0.005 | 0.040 | 0.050 | 0.050 | -0.074 | -0.300 | 0.002 | 96.80 | 81.20 |
| $\theta^{\Delta B}(X_0)$ | DML with RF | 0.532 | -0.007 | 0.046 | 0.058 | 0.058 | -0.092 | -0.119 | 0.002 | 95.20 | 80.40 |
| $\theta^{\Delta B}(X_1)$ | DML with RF | 0.501 | -0.011 | 0.049 | 0.061 | 0.062 | -0.034 | 0.079 | 0.001 | 94.80 | 80.00 |
| $\theta^{\Delta B}(X_2)$ | DML with RF | 0.423 | -0.009 | 0.040 | 0.050 | 0.051 | -0.007 | -0.075 | 0.002 | 94.80 | 80.00 |
| $\theta^{\Delta B}(X_5)$ | DML with RF | 0.288 | -0.012 | 0.050 | 0.061 | 0.062 | 0.103 | -0.129 | -0.001 | 94.80 | 78.00 |
| $\theta^{\Delta B}(X_0, X_1)$ | DML with RF | 0.645 | -0.022 | 0.053 | 0.063 | 0.067 | -0.041 | 0.113 | 0.004 | 93.60 | 80.80 |
| $\theta^{\Delta B}(X_1, X_2)$ | DML with RF | 0.502 | -0.011 | 0.049 | 0.060 | 0.061 | -0.045 | 0.087 | -0.000 | 95.20 | 78.80 |
| $\theta^{\Delta B}(X_0, X_1, X_2)$ | DML with RF | 0.655 | -0.032 | 0.056 | 0.063 | 0.070 | -0.060 | 0.104 | 0.003 | 92.40 | 78.80 |

*Note:* This table shows results for $N = 10,000$, $R = 250$ for a non-linear DGP with non-linear heterogeneous treatment effects and a covariate influenced by the moderator variable in two blocks: (1) ΔBGATE results with linear regression in the second step (2) ΔBGATE results with random forest in the second step. $\theta^{\Delta G}$ shows the difference between two GATEs. Column (1) shows the effect estimated, and column (2) shows the second estimation step. The remaining ten columns depict performance measures explained in C.1.

## C.3 Details for $\Delta$CBGATE

### C.3.1 Data Generating Process (DGP)

The DGP is nearly identical to the one used in the first simulation study for the $\Delta$BGATE. We start with simulating a p-dimensional covariate matrix $X_{i,p}$ with $p = 6$. The first two covariates are drawn from a uniform distribution $X_{i,0}, X_{i,1} \sim \mathcal{U}[0,1]$ and the remaining covariates from a normal distribution $X_{i,2} \ldots, X_{i,p-1} \sim \mathcal{N}(0.5, \sqrt{1/12})$. All covariates have a mean of 0.5 and a standard deviation of $\sqrt{1/12}$. In the simulation design that features a correlation between the moderator $Z_i$ and some of the covariates $X_i$, the moderator variable is created like in the simulation study for the $\Delta$BGATE. Hence the moderator variable $Z_i$ is drawn from a Bernoulli distribution with probability $P(Z_i = 1|X_{i,0}, X_{i,1}) = (0.1 + 0.8\beta(X_{i,0} \times X_{i,1}; 2, 4))$.[15] The propensity score is created similarly as in Künzel, Sekhon, Bickel, & Yu (2019) and Wager & Athey (2018). In the second simulation design, there is no correlation between the moderator $Z_i$ and some $X_i$. We draw noise $e_i \sim \mathcal{N}(0,1)$. If $e_i > 0.5$, the moderator variable takes the value 1, and otherwise 0. The treatment variable $D_i$ is drawn from a Bernoulli distribution with probability $P(D_i = 1|X_{i,0}, X_{i,1}, X_{i,2}, X_{i,5}, Z_i) = (0.2 + 0.6\beta(\frac{X_{i,0}+X_{i,1}+X_{i,2}+X_{i,5}+Z_i}{5}; 2, 4))$.

Next, the response functions under treatment and non-treatment and the two states of the moderator variable are specified. The non-treatment response function is specified similarly as in Nie & Wager (2021) and creates a difficult non-linear setting. They are given by

$$\mu_0(1, X_i) = \sin(\pi \times X_{i,0} \times X_{i,1}) + (X_{i,2} - 0.5)^2 + 0.1X_{i,3} + 0.3X_{i,5}$$
$$\mu_0(0, X_i) = \sin(\pi \times X_{i,0} \times X_{i,1}) + (X_{i,2} - 0.5)^2 + 0.1X_{i,3} + 0.3X_{i,5}$$

The response functions under treatment are defined differently, namely as:

$$\mu_1(1, X_i) = \mu_0(1, X_i) + \sin(4.9X_{i,0}) + \sin(2X_{i,1}) + 0.7X_{i,2}^4 + 0.4X_{i,5}$$
$$\mu_1(0, X_i) = \mu_0(0, X_i) + \sin(1.4X_{i,0}) + \sin(6X_{i,1}) + 0.6X_{i,2}^2 + 0.3X_{i,5}.$$

In contrast to the response functions in the $\Delta$BGATE simulation, they are not directly influenced by the moderator $Z_i$ because it would violate Assumption 10. Hence, it would not be possible to identify a $\Delta$CBGATE. Moreover, we restrict this simulation to non-linear heterogeneitey.

Last, we simulate the potential outcomes as $Y_i^{d,z} = \mu_d(z, X_i) + e_{i,d,z}$ with noise $e_{i,d,z} \sim \mathcal{N}(0,1)$.

[15] $\beta(X_{i,0} \times X_{i,1}; 2, 4)$ denotes the CDF of a beta distribution with the shape parameters $\alpha = 2$ and $\beta = 4$.

Summing up, the data consists of an observable quadruple $(y_{i,r}, d_{i,r}, z_{i,r}, x_{i,r})$ and the true values are estimated on a sample with $N = 100,000$.

### C.3.2 Effects of interest and estimators

To compare the ΔCBGATE with the ΔGATE, consider both effects:

$$\theta^{\Delta G} = \mathrm{E}\left[Y_i^{1,1} - Y_i^{0,1}|Z_i = 1\right] - \mathrm{E}\left[Y_i^{1,0} - Y_i^{0,0}|Z_i = 0\right] \tag{C.1}$$

$$\theta^{\Delta C} = \mathrm{E}\left[Y_i^{1,1} - Y_i^{0,1} - Y_i^{1,0} + Y_i^{0,0}\right] \tag{C.2}$$

These two effects are only identical if there is no correlation between the moderator $Z_i$ and the covariates $X_i$. We estimate Equation C.1 and C.2 with DML with 5-fold cross-fitting in the following versions:

**Table C.8:** Simulation study: Versions of the DML estimators

| | $\theta^{\Delta C}$ | $\theta^{\Delta G}$ |
|---|---|---|
| | propensity score | propensity score |
| (1) | $P(D_i = d, Z_i = z\|X_i = x)$ | $P(D_i = d\|X_i = x)$ |
| (2) | $P(D_i = d\|Z_i = z, X_i = x)$, $P(Z_i = z\|X_i = x)$ | |

*Note:* This table depicts different versions of DML estimators used for $\theta^{\Delta C}$. We use either the two marginal propensity scores or the joint propensity score. All weights have been normalized using Algorithm 2.

Random forests (number of trees: 1,000) are used to estimate the nuisance functions. Algorithm 4 in Section 3 shows the ΔCBGATE, while Algorithm 5 in Appendix C summarizes the implemented GATE estimator. The GATE is estimated by separately estimating an ATE in the groups $Z_i = 1$ and $Z_i = 0$. To obtain $\theta^{\Delta G}$, the difference between those two ATEs is taken.

### C.3.3 Simulation Design

In total, four different simulation settings are estimated. The first element we change across settings are the numbers of replications $R$ and observations $N$. We run two simulations with $N = 2,500$ and $R = 1,000$ and two with $N = 10,000$ and $R = 250$. Furthermore, as explained above, we vary the correlation of $X$ and $Z$. The optimal hyperparameters for the random forests, namely the maximum tree depth and the minimum leaf size, are again tuned by a grid search (maximum tree depth: [5, 10, 15, 20], minimum leaf size: [2, 5, 10, 20]). They are tuned using five different data draws and then fixed for the whole simulation. The optimal combination of maximal tree depth and minimum leaf size is chosen by taking the combination that appears most often in the five draws. The optimal parameters are shown in Table C.9.

**Table C.9:** Simulation Study: Optimal hyperparameters for random forests ($\Delta$CBGATE)

| | $\mu_1$ | $\mu_0$ | $\mu_{11}$ | $\mu_{01}$ | $\mu_{10}$ | $\mu_{00}$ | $\pi_d$ | $\lambda_z$ | $\omega_{11}$ | $\omega_{01}$ | $\omega_{10}$ | $\omega_{00}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP: correlation between moderator $Z_i$ and some covariates $X_i$ | | | | | | | | | | | | |
| N = 2,500 | | | | | | | | | | | | |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 10 | 5 | 5 | 5 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 2 | 5 | 2 | 2 | 10 | 20 | 2 | 20 | 20 | 20 | 10 | 20 |
| N = 10,000 | | | | | | | | | | | | |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 5 | 5 | 5 | 10 | 5 | 5 | 10 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 10 | 2 | 2 | 20 | 20 | 20 | 20 | 2 | 20 | 20 | 5 | 20 |
| DGP: no correlation between moderator $Z_i$ and some covariates $X_i$ | | | | | | | | | | | | |
| N = 2,500 | | | | | | | | | | | | |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 10 | 5 | 10 | 5 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 5 | 10 | 5 | 20 | 10 | 20 | 20 | 20 | 20 | 20 | 20 | 10 |
| N = 10,000 | | | | | | | | | | | | |
| Number of trees | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Maximal tree depth | 10 | 5 | 10 | 5 | 10 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Minimum leaf size | 5 | 10 | 10 | 20 | 20 | 5 | 2 | 10 | 20 | 20 | 20 | 20 |

*Note:* This table depicts the optimal hyperparameters for the random forests. The grid values are as follows: maximum tree depth: [5, 10, 15, 20], minimum leaf size: [2, 5, 10, 20].

### C.3.4 Results

Figures C.1 - C.4 compare the distribution of the biases for $\hat{\theta}^{\Delta C}$ and $\hat{\theta}^{\Delta G}$ to a normal distribution. The bias of both estimators is calculated using the true $\theta^{\Delta C}$. The first finding is that the $\hat{\theta}^{\Delta G}$ estimator can be substantially biased if $Z_i$ and $X_i$ are correlated. This bias can be seen in Figure C.1 and C.3. In addition, comparing the two different sample sizes shows that if $N$ increases from 2,500 to 10,000, the standard error halves. Looking at Figure C.2 and C.4, the estimators of $\hat{\theta}^{\Delta C}$ and $\hat{\theta}^{\Delta G}$ are very similar when there is no correlation between $Z_i$ and $X_i$. Moreover, their distributions appear to converge to a normal distribution.

Table C.10 shows the results for $N = 2,500$ and $R = 1,000$. The first block presents the results for the setting where there is selection into treatment and correlation between $Z_i$ and $X_i$. The second block depicts the results without correlating $Z_i$ and $X_i$. Estimating the effect ($\hat{\theta}^{\Delta C}$) with a joint propensity score (Specification 1) or with the product of two marginal propensity scores (Specification 2) leads to qualitatively similar results.

**Figure C.1:** N = 2,500, correlated $Z_i$ and $X_i$

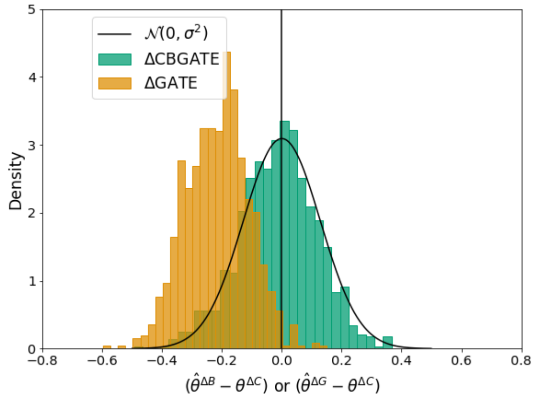



**Figure C.2:** N = 2500, uncorrelated $Z_i$ and $X_i$

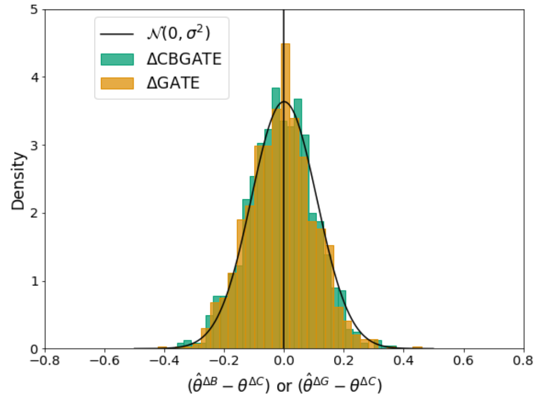



**Figure C.3:** N = 10,000, correlated $Z_i$ and $X_i$

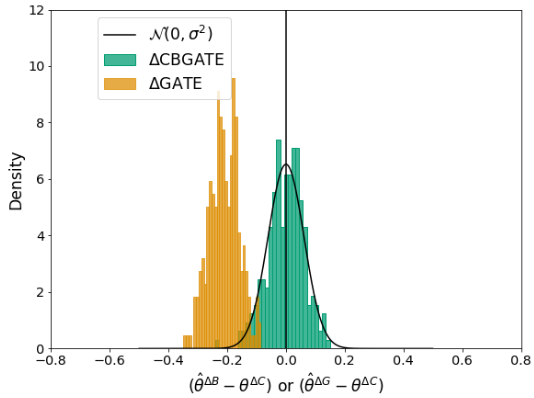
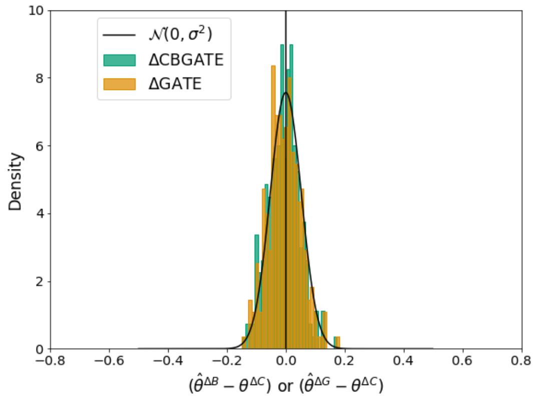



**Figure C.4:** N = 10,000, uncorrelated $Z_i$ and $X_i$




*Note:* These figures show the distribution of the bias of $\hat{\theta}^{\Delta C}$ and $\hat{\theta}^{\Delta G}$ compared to the true $\theta^{\Delta C}$. We plot a normal distribution to indicate convergence to an appropriately normal distribution. The figures are created using the results of the unconfoundedness setting and the DML estimator with the joint propensity score estimator.

Table C.11 shows results for $N = 10,000$ and $R = 250$. Compared to the results with the smaller sample presented in Table C.10, the standard errors and the RMSE halve. This aligns with the convergence rate of a $\sqrt{N}$-consistent estimator.

# D Appendix: Empirical Example

## D.1 Data Descriptives

Table D.1 shows the mean and standard deviation of the covariates included in the analyzes by treatment and moderator status. In addition to the covariates in this table, we add caseworkers' fixed effects. For a more detailed description of the data and how the dataset was constructed, please see Knaus et al. (2022). The data can be accessed on swissubase.ch for research purposes. Table D.2 shows the mean and standard deviation of the covariates included in the analyzes by treatment status and Table D.3 by moderator status.

**Table C.10:** Simulation results: N = 2,500

| | Number of observations = 2,500 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimation of effects | | | | | | | Estimation of standard errors | | | |
| | Spf. | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| | **Correlation between $Z_i$ and $X_i$** | | | | | | | | | | |
| $\theta^{\Delta G}$ | 1 | 0.072 | -0.012 | 0.084 | 0.104 | 0.105 | -0.017 | -0.053 | -0.000 | 95.30 | 79.50 |
| $\theta^{\Delta C}$ | 1 | 0.277 | -0.013 | 0.102 | 0.129 | 0.130 | -0.128 | 0.137 | -0.001 | 94.00 | 80.40 |
| $\theta^{\Delta C}$ | 2 | 0.277 | -0.019 | 0.101 | 0.126 | 0.128 | -0.150 | 0.053 | 0.001 | 94.30 | 80.30 |
| | **No correlation between $Z_i$ and $X_i$** | | | | | | | | | | |
| $\theta^{\Delta G}$ | 1 | 0.278 | -0.014 | 0.086 | 0.109 | 0.110 | 0.049 | 0.324 | -0.000 | 94.10 | 79.50 |
| $\theta^{\Delta C}$ | 1 | 0.277 | -0.011 | 0.088 | 0.110 | 0.110 | -0.048 | 0.087 | -0.001 | 94.30 | 79.80 |
| $\theta^{\Delta C}$ | 2 | 0.277 | -0.017 | 0.087 | 0.109 | 0.110 | -0.028 | 0.111 | -0.001 | 94.50 | 79.20 |

*Note:* This table shows results for $N = 2,500$ and $R = 1,000$ in two blocks: (1) correlation between $Z_i$ and $X_i$, and (2) no correlation between $Z_i$ and $X_i$. Column (1) shows the effect estimated, column (2) the respective version of the DML estimator used as explained in Table C.8 (3) the true effect. The remaining ten columns depict performance measures explained in C.1.

**Table C.11:** Simulation results: N = 10,000

| | Number of observations = 10,000 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimation of effects | | | | | | | Estimation of standard errors | | | |
| | Spf. | Truth | Bias | \|Bias\| | Std | RMSE | Skew. | Ex. Kurt. | Bias (SE) | CovP (95) (in %) | CovP (80) (in %) |
| | **Correlation between $Z_i$ and $X_i$** | | | | | | | | | | |
| $\theta^{\Delta G}$ | 1 | 0.072 | -0.007 | 0.041 | 0.050 | 0.050 | -0.033 | -0.258 | 0.000 | 94.40 | 80.40 |
| $\theta^{\Delta C}$ | 1 | 0.277 | -0.002 | 0.049 | 0.061 | 0.061 | -0.338 | 0.349 | 0.002 | 95.60 | 80.80 |
| $\theta^{\Delta C}$ | 2 | 0.277 | 0.001 | 0.050 | 0.062 | 0.062 | -0.426 | 0.523 | 0.000 | 95.20 | 80.00 |
| | **No correlation between $Z_i$ and $X_i$** | | | | | | | | | | |
| $\theta^{\Delta G}$ | 1 | 0.278 | -0.008 | 0.045 | 0.055 | 0.055 | 0.199 | 0.064 | -0.002 | 93.60 | 77.60 |
| $\theta^{\Delta C}$ | 1 | 0.277 | -0.009 | 0.042 | 0.053 | 0.053 | 0.093 | 0.170 | -0.001 | 94.00 | 78.00 |
| $\theta^{\Delta C}$ | 2 | 0.277 | -0.010 | 0.043 | 0.053 | 0.054 | 0.114 | 0.217 | -0.002 | 93.60 | 78.00 |

*Note:* This table shows results for $N = 10,500$ and $R = 250$ in two blocks: (1) correlation between $Z_i$ and $X_i$, and (2) no correlation between $Z_i$ and $X_i$. Column (1) shows the effect estimated, column (2) the respective version of the DML estimator used as explained in Table C.8 (3) the true effect. The remaining ten columns depict performance measures explained in C.1.

**Table D.1:** Empirical analysis: Balance table for treatment and moderator variable

| Variable | **Treated Non-Swiss** Mean | **Treated Swiss** Mean | **Non-treated Non-Swiss** Mean | **Non-treated Swiss** Mean |
|---|---|---|---|---|
| Age | 35.62 | 38.08 | 36.37 | 36.50 |
| French speaking canton | 0.10 | 0.07 | 0.28 | 0.23 |
| German speaking canton | 0.86 | 0.91 | 0.62 | 0.70 |
| Italian speaking canton | 0.04 | 0.02 | 0.10 | 0.07 |
| Lives in big city | 0.21 | 0.14 | 0.19 | 0.15 |
| Lives in medium city | 0.17 | 0.15 | 0.15 | 0.12 |
| Lives in contryside | 0.62 | 0.71 | 0.67 | 0.73 |
| Age of caseworker | 43.70 | 44.54 | 44.23 | 44.44 |
| Caseworkers cooperation | 0.51 | 0.48 | 0.50 | 0.47 |
| Caseworkers education: above vocational training | 0.45 | 0.45 | 0.43 | 0.44 |
| Caseworkers education: tertiary level | 0.20 | 0.22 | 0.25 | 0.23 |
| Caseworker female | 0.42 | 0.47 | 0.38 | 0.42 |
| Caseworker missing | 0.04 | 0.04 | 0.04 | 0.04 |
| Caseworker own unemployment experience | 0.63 | 0.63 | 0.63 | 0.63 |
| Caseworker job tenure | 5.53 | 5.55 | 5.90 | 5.84 |
| Caseworker education: vocational degree | 0.25 | 0.27 | 0.22 | 0.23 |
| Fraction months employed last 2 years | 0.82 | 0.84 | 0.77 | 0.81 |
| No employment spells last 5 years | 1.05 | 0.97 | 1.40 | 1.21 |
| Employability | 1.94 | 2.00 | 1.95 | 2.01 |
| Female | 0.40 | 0.47 | 0.41 | 0.46 |
| Cantonal GDP per capita | 0.51 | 0.51 | 0.49 | 0.49 |
| Married | 0.67 | 0.36 | 0.70 | 0.35 |
| Mother tongue not Swiss language | 0.65 | 0.10 | 0.66 | 0.10 |
| Past annual income | 41704 | 47899 | 38226 | 43865 |
| Previous job: manager | 0.05 | 0.09 | 0.04 | 0.09 |
| Previous job: primary sector | 0.07 | 0.05 | 0.13 | 0.08 |
| Previous job: secondary sector | 0.18 | 0.15 | 0.13 | 0.13 |
| Previous job: tertiary sector | 0.54 | 0.67 | 0.48 | 0.64 |
| Previous job: missing sector | 0.21 | 0.13 | 0.26 | 0.15 |
| Previous job: self-employed | 0.00 | 0.00 | 0.01 | 0.01 |
| Previous job: skilled worker | 0.47 | 0.72 | 0.45 | 0.70 |
| Previous job: unskilled worker | 0.47 | 0.16 | 0.48 | 0.17 |
| Qualification: some degree | 0.35 | 0.73 | 0.32 | 0.73 |
| Qualification: semiskilled | 0.18 | 0.13 | 0.21 | 0.13 |
| Qualification: unskilled | 0.40 | 0.13 | 0.40 | 0.12 |
| Qualification: skilled without degree | 0.07 | 0.02 | 0.08 | 0.03 |
| Allocation to caseworker: by industry | 0.63 | 0.67 | 0.53 | 0.54 |
| Allocation to caseworker: by occupation | 0.56 | 0.59 | 0.56 | 0.56 |
| Allocation to caseworker: by age | 0.04 | 0.04 | 0.03 | 0.03 |
| Allocation to caseworker: by employability | 0.07 | 0.07 | 0.06 | 0.07 |
| Allocation to caseworker: by region | 0.09 | 0.09 | 0.13 | 0.12 |
| Allocation to caseworker: other | 0.06 | 0.07 | 0.07 | 0.08 |
| No of unemployment spells last 2 years | 0.54 | 0.35 | 0.78 | 0.50 |
| Cantonal unemployment rate | 3.71 | 3.60 | 3.82 | 3.69 |
| Number of observations | 4,438 | 8,607 | 30,417 | 47,877 |

*Note:* This table shows the mean of some covariates included in the analysis. Column (1) and (2) show it for treated individuals, column (3) and (4) for non-treated individuals. Column (1) and (3) show it for non-Swiss individuals, column (2) and (4) for Swiss individuals.

**Table D.2:** Empirical analysis: Balance table for treatment variable (participation in the program)

| | **Treated** | | **Control** | | |
|---|---|---|---|---|---|
| Covariates | Mean | Std. Dev. | Mean | Std. Dev. | Std. Diff. |
| Age | 37.24 | 8.78 | 36.45 | 8.64 | 9.13 |
| French speaking canton | 0.08 | 0.27 | 0.25 | 0.43 | 46.44 |
| German speaking canton | 0.89 | 0.31 | 0.67 | 0.47 | 55.63 |
| Italian speaking canton | 0.03 | 0.16 | 0.08 | 0.27 | 24.01 |
| Mother tongue in cantonal language | 0.12 | 0.33 | 0.11 | 0.32 | 3.71 |
| Lives in big city | 0.17 | 0.37 | 0.16 | 0.37 | 0.54 |
| Lives in medium city | 0.16 | 0.36 | 0.13 | 0.34 | 7.39 |
| Lives in contryside | 0.68 | 0.47 | 0.71 | 0.46 | 6.04 |
| Age of caseworker | 44.26 | 11.64 | 44.36 | 11.60 | 0.88 |
| Caseworkers cooperation | 0.49 | 0.50 | 0.48 | 0.50 | 1.73 |
| Caseworkers education: above vocational training | 0.45 | 0.50 | 0.43 | 0.50 | 3.38 |
| Caseworkers education: tertiary level | 0.21 | 0.41 | 0.24 | 0.43 | 6.36 |
| Caseworker female | 0.45 | 0.50 | 0.41 | 0.49 | 9.98 |
| Caseworker missing | 0.04 | 0.20 | 0.04 | 0.20 | 0.09 |
| Caseworker own unemployment experience | 0.63 | 0.48 | 0.63 | 0.48 | 1.01 |
| Caseworker job tenure | 5.54 | 3.23 | 5.86 | 3.30 | 9.73 |
| Caseworker education: vocational degree | 0.26 | 0.44 | 0.23 | 0.42 | 7.80 |
| Fraction months employed last 2 years | 0.83 | 0.22 | 0.79 | 0.25 | 16.86 |
| No employment spells last 5 years | 1.00 | 1.27 | 1.28 | 1.52 | 20.44 |
| Employability | 1.98 | 0.48 | 1.99 | 0.51 | 1.10 |
| Female | 0.44 | 0.50 | 0.44 | 0.50 | 1.33 |
| Foreigner with permit B | 0.11 | 0.31 | 0.14 | 0.35 | 9.68 |
| Foreigner with permit C | 0.23 | 0.42 | 0.25 | 0.43 | 3.87 |
| Cantonal GDP per capita | 0.51 | 0.09 | 0.49 | 0.09 | 22.15 |
| Married | 0.47 | 0.50 | 0.48 | 0.50 | 2.66 |
| Mother tongue not Swiss language | 0.29 | 0.45 | 0.32 | 0.47 | 6.82 |
| Past annual income | 45792 | 20184 | 41674 | 20459 | 20 |
| Previous job: manager | 0.08 | 0.27 | 0.07 | 0.26 | 3.01 |
| Previous job: primary sector | 0.06 | 0.23 | 0.10 | 0.30 | 15.25 |
| Previous job: secondary sector | 0.16 | 0.37 | 0.13 | 0.34 | 8.72 |
| Previous job: tertiary sector | 0.63 | 0.48 | 0.58 | 0.49 | 10.42 |
| Previous job: missing sector | 0.15 | 0.36 | 0.19 | 0.40 | 10.74 |
| Previous job: self-employed | 0.00 | 0.06 | 0.01 | 0.08 | 4.15 |
| Previous job: skilled worker | 0.63 | 0.48 | 0.61 | 0.49 | 5.46 |
| Previous job: unskilled worker | 0.26 | 0.44 | 0.29 | 0.45 | 5.99 |
| Qualification: some degree | 0.60 | 0.49 | 0.57 | 0.50 | 6.09 |
| Qualification: semiskilled | 0.15 | 0.35 | 0.16 | 0.37 | 3.31 |
| Qualification: unskilled | 0.22 | 0.41 | 0.23 | 0.42 | 1.41 |
| Qualification: skilled without degree | 0.03 | 0.18 | 0.05 | 0.21 | 6.28 |
| Swiss | 0.66 | 0.47 | 0.61 | 0.49 | 10.05 |
| Allocation to caseworker: by industry | 0.66 | 0.47 | 0.53 | 0.50 | 25.25 |
| Allocation to caseworker: by occupation | 0.58 | 0.49 | 0.56 | 0.50 | 4.64 |
| Allocation to caseworker: by age | 0.04 | 0.19 | 0.03 | 0.17 | 3.53 |
| Allocation to caseworker: by employability | 0.07 | 0.25 | 0.07 | 0.25 | 0.02 |
| Allocation to caseworker: by region | 0.09 | 0.28 | 0.12 | 0.33 | 10.93 |
| Allocation to caseworker: other | 0.07 | 0.25 | 0.07 | 0.26 | 1.97 |
| No of unemployment spells last 2 years | 0.41 | 1.00 | 0.61 | 1.26 | 17.55 |
| Cantonal unemployment rate | 3.64 | 0.77 | 3.74 | 0.86 | 12.45 |
| Number of observations | 13,045 | | 78,294 | | |

*Note:* This table shows the mean and standard deviation of the covariates included in the analyzes. Columns (1) and (2) show it for the treated group, columns (3) and (4) for the control group. The last column shows the standardized difference between the two groups. The standardized difference is calculated as $SD = \frac{|\bar{X}_{\text{treated}} - \bar{X}_{\text{control}}|}{\sqrt{1/2\left(\text{Var}(\bar{X}_{\text{treated}}) + \text{Var}(\bar{X}_{\text{control}})\right)}} \cdot 100$ where $\bar{X}_{\text{treated}}$ and $\bar{X}_{\text{control}}$ indicate the sample mean of the treatment and control group, respectively.

**Table D.3:** Empirical analysis: Balance table for moderator variable (nationality)

| | **Non Swiss** | | **Swiss** | | |
|---|---|---|---|---|---|
| Covariates | Mean | Std. Dev. | Mean | Std. Dev. | Std. Diff. |
| Age | 36.27 | 8.22 | 36.74 | 8.92 | 5.38 |
| French speaking canton | 0.25 | 0.44 | 0.20 | 0.40 | 11.92 |
| German speaking canton | 0.65 | 0.48 | 0.74 | 0.44 | 17.80 |
| Italian speaking canton | 0.09 | 0.29 | 0.06 | 0.24 | 11.96 |
| Lives in big city | 0.19 | 0.39 | 0.15 | 0.36 | 10.47 |
| Lives in medium city | 0.15 | 0.36 | 0.13 | 0.33 | 7.01 |
| Lives in contryside | 0.66 | 0.47 | 0.73 | 0.45 | 13.76 |
| Age of caseworker | 44.16 | 11.69 | 44.46 | 11.55 | 2.53 |
| Caseworkers cooperation | 0.50 | 0.50 | 0.47 | 0.50 | 6.45 |
| Caseworkers education: above vocational training | 0.43 | 0.49 | 0.44 | 0.50 | 2.83 |
| Caseworkers education: tertiary level | 0.24 | 0.43 | 0.23 | 0.42 | 2.45 |
| Caseworker female | 0.39 | 0.49 | 0.43 | 0.49 | 7.89 |
| Caseworker missing | 0.04 | 0.21 | 0.04 | 0.20 | 2.30 |
| Caseworker own unemployment experience | 0.63 | 0.48 | 0.63 | 0.48 | 0.35 |
| Caseworker job tenure | 5.85 | 3.36 | 5.80 | 3.26 | 1.56 |
| Caseworker education: vocational degree | 0.23 | 0.42 | 0.23 | 0.42 | 1.91 |
| Fraction months employed last 2 years | 0.78 | 0.26 | 0.81 | 0.24 | 15.26 |
| No employment spells last 5 years | 1.36 | 1.62 | 1.17 | 1.40 | 12.54 |
| Employability | 1.95 | 0.51 | 2.01 | 0.50 | 12.50 |
| Female | 0.41 | 0.49 | 0.46 | 0.50 | 9.98 |
| Cantonal GDP per capita | 0.49 | 0.09 | 0.50 | 0.09 | 2.97 |
| Married | 0.69 | 0.46 | 0.35 | 0.48 | 73.50 |
| Mother tongue not Swiss language | 0.66 | 0.47 | 0.10 | 0.30 | 141.73 |
| Past annual income | 38669 | 18536 | 44480 | 21278 | 29 |
| Previous job: manager | 0.05 | 0.21 | 0.09 | 0.29 | 17.74 |
| Previous job: primary sector | 0.12 | 0.33 | 0.07 | 0.26 | 16.09 |
| Previous job: secondary sector | 0.14 | 0.35 | 0.13 | 0.34 | 2.53 |
| Previous job: tertiary sector | 0.48 | 0.50 | 0.65 | 0.48 | 33.52 |
| Previous job: missing sector | 0.25 | 0.44 | 0.15 | 0.35 | 27.04 |
| Previous job: self-employed | 0.00 | 0.07 | 0.01 | 0.08 | 2.59 |
| Previous job: skilled worker | 0.46 | 0.50 | 0.70 | 0.46 | 51.92 |
| Previous job: unskilled worker | 0.48 | 0.50 | 0.16 | 0.37 | 71.85 |
| Qualification: some degree | 0.32 | 0.47 | 0.73 | 0.44 | 89.33 |
| Qualification: semiskilled | 0.21 | 0.40 | 0.13 | 0.33 | 21.16 |
| Qualification: unskilled | 0.40 | 0.49 | 0.12 | 0.32 | 66.90 |
| Qualification: skilled without degree | 0.08 | 0.26 | 0.02 | 0.15 | 23.58 |
| Allocation to caseworker: by industry | 0.54 | 0.50 | 0.56 | 0.50 | 3.80 |
| Allocation to caseworker: by occupation | 0.56 | 0.50 | 0.57 | 0.50 | 1.97 |
| Allocation to caseworker: by age | 0.03 | 0.17 | 0.03 | 0.18 | 2.80 |
| Allocation to caseworker: by employability | 0.06 | 0.24 | 0.07 | 0.26 | 5.08 |
| Allocation to caseworker: by region | 0.12 | 0.33 | 0.12 | 0.32 | 1.52 |
| Allocation to caseworker: other | 0.07 | 0.26 | 0.08 | 0.27 | 2.42 |
| No of unemployment spells last 2 years | 0.75 | 1.38 | 0.48 | 1.10 | 21.73 |
| Cantonal unemployment rate | 3.81 | 0.83 | 3.68 | 0.85 | 15.33 |
| Number of observations | 34,855 | | 56,484 | | |

*Note:* This table shows the mean and standard deviation of some covariates included in the analysis. Column (1) and (2) show it for Swiss individuals, column (3) and (4) for non-Swiss individuals. The last column shows the standardized difference between the two groups. The standardized difference is calculated as $SD = \frac{|\bar{X}_{\text{Swiss}} - \bar{X}_{\text{non-Swiss}}|}{\sqrt{1/2\left(\text{Var}\left(\bar{X}_{\text{Swiss}}\right)+\text{Var}\left(\bar{X}_{\text{non-Swiss}}\right)\right)}} \cdot 100$ where $\bar{X}_{\text{Swiss}}$ and $\bar{X}_{\text{non-Swiss}}$ indicate the sample mean of the Non-Swiss and the Swiss individuals, respectively.