Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments

Phillip Heiler[†] Michael C. Knaus[‡]

First version: October 04, 2021
This version: September 23, 2022

Abstract

Binary treatments are often ex-post aggregates of multiple treatments or can be disaggregated into multiple treatment versions. Thus, effects can be heterogeneous due to either effect or treatment heterogeneity. We propose a decomposition method that uncovers masked heterogeneity, avoids spurious discoveries, and evaluates treatment assignment quality. The estimation and inference procedure based on double/debiased machine learning allows for high-dimensional confounding, many treatments and extreme propensity scores. Our applications suggest that heterogeneous effects of smoking on birthweight are partially due to different smoking intensities and that gender gaps in Job Corps effectiveness are largely explained by differences in vocational training.

Keywords: causal inference, causal machine learning, double machine learning, heterogeneous treatment effects, overlap, treatment versions

JEL classification: C14, C21

*Michael Knaus gratefully acknowledges financial support from the Swiss National Science Foundation (SNSF) (grant number SNSF 407740_187301). The paper was circulated and presented previously under different titles. We would like to thank Martin Huber, Michael Lechner, Jana Mareckova, Julian Schüssler, Anthony Strittmatter, and the participants of sessions at IAAE2021, ESEM2021, COMPIE2021, and the FFHKT Econometrics Seminar for valuable comments and discussions. All remaining errors are ours.

[†]Aarhus University, Department of Economics and Business Economics, CREATES, TrygFonden's Centre for Child Research, Fuglesangs Allé 4, 8210 Aarhus V, Denmark, pheiler@econ.au.dk.

[‡]University of Tübingen, Mohlstraße. 36, 72074 Tübingen, Germany. Michael C. Knaus is also affiliated with IZA, Bonn, michael.knaus@uni-tuebingen.de.

1 Introduction

The analysis of causal effects is at the heart of empirical research in economics, political science, the biomedical sciences, and beyond. To evaluate and design policies, interventions, or programs for observational units with different background characteristics, it is crucial to develop a thorough understanding of the heterogeneity present in causal relationships. There is now a large literature that develops and applies identification and estimation strategies for causal or treatment parameters that explicitly take into account such heterogeneity, see Athey and Imbens (2017) or Abadie and Cattaneo (2018) for recent overviews.

Most attention is on *effect heterogeneity* of binary treatments, while less is given to *treatment heterogeneity*. However, many binary treatments in applications can be conceived as heterogeneous in the sense that they summarize (many) underlying *effective treatments* that impact the outcome of interest. In such cases it is not clear whether effect heterogeneity as defined in the canonical binary treatment setting reflects heterogeneous effects or heterogeneity in the effective treatments. This paper proposes new estimands and estimators to disentangle these sources of heterogeneity in a general setting where the analyzed binary indicator does not coincide with the effective treatment. The distinction between sources of heterogeneity is crucial for evaluating and improving assignment mechanisms. Consider the following two scenarios:[1]

Scenario 1 (binarized treatments): Multiple or continuous treatments are *ex-post* subsumed into a binary indicator (e.g. different smoking intensities become "smoking yes/no"). Such aggregations are often motivated by simplicity or data availability, but can have unintentional consequences: First, discovered effect heterogeneity can be a spurious byproduct of aggregation and thus falsely be attributed to unit background characteristics. Second, actual effect heterogeneity could be masked as a consequence of the aggregation.

Scenario 2 (multiple treatment versions): A binary treatment takes different versions after assignment, e.g. access to a training program with different specializations. Here, effect heterogeneity could result from better version targeting and not from different effectiveness of the versions themselves. Understanding this difference is crucial for policy

[1]See also Supplementary Appendix B.1 for a motivating toy example.

makers to assess the quality of the version assignment mechanism.

In this paper we propose a novel method for decomposing effect heterogeneity in a more general scenario with observed confounders. We decompose canonical effect heterogeneity into new estimands that are representative of (i) heterogeneous effects and (ii) heterogeneity stemming from different underlying treatment compositions. These decomposition parameters serve as summary measures to evaluate the consequences of (dis)aggregating treatment variables for the causal analysis. Furthermore they provide a simple framework for comparing the quality of treatment version assignment rules and their heterogeneity across units or groups.

We develop a simple but flexible nonparametric method for estimation and statistical inference for the decomposition parameters. Our framework allows for the use of machine learning techniques such as random forests, deep neural networks, high-dimensional sparse regression models in the estimation of the nuisance parameters. We provide explicit high-level conditions regarding the required rates for machine learners, their interaction with the nonparametric decomposition step, and the number of effective treatments J. We also provide sufficient conditions for explicit examples.

The decomposition can be used to conduct simple joint hypothesis tests for global or conditional decomposition parameters that consider all effective treatments simultaneously. This allows to test necessary conditions for different types of selection and effect heterogeneity without the need for multiple testing procedures. It compares favorably to conventional multi-valued treatment effect analysis under many effective treatments $J \to \infty$, expanding sets of nuisance parameters, and extreme propensity scores. In particular, regular inference is still achievable even if propensity scores are arbitrarily close to zero (limited overlap). This result is obtained by leveraging superefficiency properties of (conditional) probability estimators.

The large sample theory also extends beyond the decomposition parameters considered in this paper. In particular, we provide conditions under which unbiased signals with machine learning inputs for causal parameters can be combined with weighting schemes that are themselves estimated to obtain asymptotically valid analytical confidence intervals. Our Monte Carlo simulations suggest that the proposed intervals have coverage rates close

to the nominal level in finite samples.

We provide two applications of our decomposition method, one for each leading scenario: First, we show that parts of the finding that the detrimental effect of smoking on birth weight is largest for white mothers can be explained by white mothers smoking more heavily conditional on being smokers. Similarly, different effects for different age groups are partly due to teenage mothers smoking less intensively compared to older mothers. Second, we investigate the lower effectiveness of access to the Job Corps training program for women compared to men. We find evidence that the gender gap is largely explained by the vocational training curriculum, which focuses more on lower paying service jobs for women and more on higher paying craft jobs for men. Imposing the same mix of vocational training as part of our decomposition removes 73% of the total gender differences in the effect on earnings.

The paper is structured as follows: Section 2 discusses the related literature. Section 3 outlines the decomposition of the causal effect parameters and discusses their identification. Section 4 contains the estimation and inference method. Section 5 introduces the technical assumptions and discusses the large sample properties. Section 6 provides the Monte Carlo study. Section 7 contains the application. Section 8 concludes. We also provide an implementation in R and replication notebooks.[2]

2 Related Literature

The proposed decomposition complements the literature that considers (dis)aggregated binary treatments. Lechner (2002) discusses how to aggregate average effects of multiple treatments into composite treatment effects. Hotz, Imbens, and Mortimer (2005) and Hotz, Imbens, and Klerman (2006) discuss summarizing different training components in one binary indicator and emphasize the potential lack of external validity under latent treatment heterogeneity. Similarly, a recent stream of papers formalizes structural causal models and interpretations of compound treatments (Cole & Frangakis, 2009; VanderWeele, 2009; Hernán & VanderWeele, 2011; Petersen, 2011). VanderWeele and Hernan (2013)

[2]For Section 7.1 see mcknaus.github.io/assets/code/Replication_NB_smoking.nb.html and for Section 7.2 mcknaus.github.io/assets/code/Replication_NB_JC.nb.html on GitHub.

note that non-homogeneous treatments violate the second component of the "Stable Unit Treatment Value Assumption" (SUTVA, Rubin, 1980): no-multiple-versions-of-treatment, which requires a homogeneous treatment or at least the treatment variation irrelevance assumption of VanderWeele (2009). VanderWeele and Hernan (2013) formalize a setting where this assumption is violated and provide several new identification results and estimands. Aggregating heterogeneous treatments has also been discussed in the context of instrumental variables (Angrist & Imbens, 1995; Marshall, 2016; Andresen & Huber, 2021), regression discontinuity designs (Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2016), and models with spillovers and interactions (Vazquez-Bare, 2022). These papers mostly discuss the consequences of (dis)aggregation of treatments on unconditional estimands and their connection to (weighted) causal effects. Our paper focuses on the consequences of (dis)aggregation on effect heterogeneity.

The focus on effect heterogeneity is motivated by the surging literature that develops (e.g. Athey & Imbens, 2016; Wager & Athey, 2018; Athey, Tibshirani, & Wager, 2019; Künzel, Sekhon, Bickel, & Yu, 2019; Knaus, Lechner, & Strittmatter, 2021; Nie & Wager, 2021) and applies (e.g. Davis & Heller, 2020; Knaus, Lechner, & Strittmatter, 2022; Buhl-Wiggers, Kerwin, Muñoz, Smith, & Thornton, 2022) flexible machine learning methods to the estimation of heterogeneous causal effects. We build on the double/debiased machine learning framework by Chernozhukov et al. (2018). They use Neyman-orthogonal score functions and sample splitting in conjunction with machine learning methods for estimation of low-dimensional parameters that depend on nuisance quantities.

Regarding heterogeneity analysis, there is now a series of papers that obtain (functional) parameters by localizing these score functions using (nonparametric) regression or machine learning methods (Lee, Okui, & Whang, 2017; Zimmert & Lechner, 2019; Fan, Hsu, Lieli, & Zhang, 2022; Semenova & Chernozhukov, 2021; Kennedy, 2020; Curth & van der Schaar, 2021; Knaus, 2022; Heiler, 2022). Our theoretical contribution builds on the structural function approach by Semenova and Chernozhukov (2021) with least squares series estimation (Newey, 1997; Belloni, Chernozhukov, Chetverikov, & Kato, 2015; Cattaneo, Farrell, & Feng, 2020). We extend some of the inferential results by Semenova and Chernozhukov (2021) to settings where pseudo-outcomes are constructed as a weighted

average of Neyman-orthogonal scores with (estimated) weights and potentially many treatments.

The paper is also related to the literature regarding inference on effect parameters under extreme propensity scores or "limited overlap" (Khan & Tamer, 2010; Rothe, 2017; Ma & Wang, 2020; Hong, Leung, & Li, 2020; Heiler & Kazak, 2021). Limited overlap occurs by construction when allowing for "many treatments" $J \to \infty$. In this case, the set of nuisance parameters is expanding and classic multi-valued treatment effect parameters (e.g. Cattaneo, 2010) are irregularly identified which complicates inference. The decomposition method, however, always yields three aggregate (functional) parameters independently of J. As a consequence, regular estimation and inference regarding heterogeneity is still feasible as long as J does not grow too fast. In finite samples, determining what constitutes a many treatments setup is difficult as J is always a finite number and a small lower bound for propensities are hard to distinguish from a zero lower bound (Rothe, 2017). Thus, a method that is robust to a potentially large number of treatments provides safeguard for empirical practice.

Many studies document the detrimental effect of smoking during pregnancy on birth weight (e.g. Almond, Chay, & Lee, 2005; Abrevaya, 2006; Cattaneo, 2010; Almond & Currie, 2011). Our methodology allows us to understand how much of the heterogeneous effects of this binarized treatment is spuriously attributed to subgroup characteristics instead of to different intensities of smoking across these subgroups.

Previous studies evaluate the US training program Job Corps from different angles based on a large scale experiment (e.g. Schochet, Burghardt, & Glazerman, 2001; Schochet, Burghardt, & McConnell, 2008; Flores, Flores-Lagunes, Gonzalez, & Neumann, 2012; Eren & Ozbelik, 2014; Strittmatter, 2019). While most of them document heterogeneous effects, we explicitly disentangle how much of the effects and their heterogeneity is driven by selection into different curricula. This provides a complementary perspective on the quality of the existing assignment mechanism.

3 Decomposition and Identification

3.1 The Setting

Assume we observe independent data (Y_i, D_i, T_i, X_i) for $i = 1, \ldots, n$. Y_i denotes the outcome of interest, $D_i \in \{0, 1\}$ is the analyzed binary indicator, $T_i \in \mathcal{T} = \{0, 1, \ldots, J\}$ indicates the effective treatment[3], and X_i contains confounding variables. We consider settings that are characterized by two features: (i) Not D_i, but the effective treatment T_i has a direct influence on the outcome creating potential outcomes $Y_i(t)$ for each $t \in \mathcal{T}$. Thus, we assume SUTVA with respect to the effective treatment such that $Y_i = \sum_t \mathbb{1}(T_i = t) Y_i(t)$. (ii) Conditional on T_i, the binary indicator D_i is deterministic, i.e. it perfectly separates the support $\mathcal{T}$. We use directed acyclic graphs (DAGs) (see e.g. Pearl, 1995) to outline our main scenarios in this setting.

Figure 1: Analyzed indicator is ex-post aggregate of confounded multiple treatment:

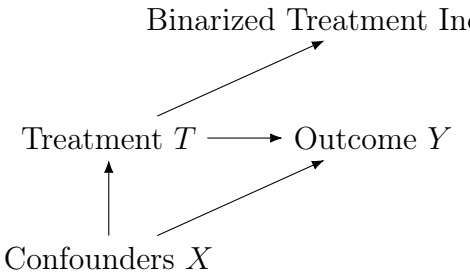
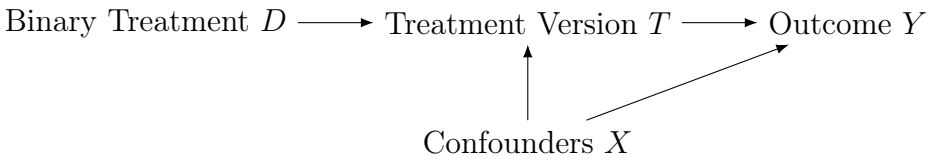


Figure 1 outlines the causal structure of *Scenario 1* where the binary indicator variable D_i is the result of an ex-post aggregation and not structurally related to the outcome. In practice, this aggregation is often conducted after the outcome realizes, which makes it unlikely for D_i to affect Y_i directly. This is indicated by a missing arrow from D_i to Y_i. However, as T_i is ancestor of both binarized indicator D_i and outcome Y_i, they are not statistically independent of each other even conditional on X_i. For example, a statistical relationship between birth weight (Y_i) and smoking (D_i) as an aggregate of the consumed dose of cigarettes (T_i) can be derived from observational data. Conditional on the number of cigarettes, however, smoking is deterministic and no association remains.

The DAG in Figure 2 depicts the causal structure of *Scenario 2* where a randomized binary treatment D_i precedes the confounded allocation of treatment versions T_i. Here,

[3]In the following the term "treatment" refers to effective treatment if not stated differently.

Figure 2: Randomized binary treatment precedes confounded treatment versions:



D_i is not an ex-post variable with regards to Y_i. Y_i and D_i are associated as the latter determines which treatment versions are available, but has no direct effect beyond that. Its effect is completely mediated through the treatment versions T_i. For example, this implies that any association between D_i being access to a training program (yes/no) on earnings Y_i would disappear if we would condition on all training types including no training access (T_i).

It is important to note that, while conceptually different in terms of the causal interpretation, the DAGs in Figures 1 and 2 imply the same conditional independence relationships regarding D_i, T_i, and potential outcomes $Y_i(t)$. In standard Neyman-Rubin notation for multi-valued treatments (Rubin, 1974; Imbens, 2000; Lechner, 2001; Cattaneo, 2010), we have that

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp D_i \mid T_i \tag{1}$$

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp T_i \mid X_i \tag{2}$$

where (1) is a consequence of our setting that D_i is a constant given T_i and (2) follows from the causal structure encoded in the DAG (see Appendix B.2.1). Thus, from a statistical point of view, we treat both scenarios as being equivalent in the following. Note that conditions (1) and (2) and everything that follows can be extended to apply to causal graphs where additional observed confounders can affect D_i, T_i, and Y_i simultaneously (see Appendix B.2.2).

We use $D_{t,i} = \mathbb{1}(T_i = t)$ to indicate that unit i is observed in treatment t and let $e_t(x) = P(D_{t,i} = 1 | X_i = x)$ denote the corresponding propensity scores. Without loss of generality, we assume throughout that $T_i = 0$ denotes a homogeneous control condition. Thus, the binary indicator is defined as $D_i = \sum_{t \neq 0} D_{t,i}$ and $D_{0,i} = 1 - D_i$ in what follows.

3.2 Heterogeneous effects if treatment heterogeneity is ignored

We are interested in the case where the causal structure is accurately described by Section 3.1, but the analyst considers only the binary indicator D_i, which is deterministic in T_i. Then, the main quantities of interest are often conditional average treatment effects ($CATE$) $\tau(x)$ or aggregations thereof like the average treatment effect ($ATE = E[\tau(X_i)]$). However, the potential outcome under the binary indicator being one is not uniquely defined in our setting unless $J = 1$. The question is then, what does the quantity $\tau(x) = E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]$ commonly deployed under strong ignorability assumptions for D_i (Imbens & Rubin, 2015) actually identify? Given the setting outlined in Section 3.1, we can backwards engineer the actually identified estimand in terms of potential outcomes of the effective treatment:

$$
\begin{aligned}
\tau(x) &= E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] \\
&= \sum_{t \neq 0} E\left[D_{t,i} Y_i(t) | D_i = 1, X_i = x\right] - E[Y_i(0)|X_i = x] \\
&= \sum_{t \neq 0} E[Y_i(t)|D_{t,i} = 1, D_i = 1, X_i = x] P(D_{t,i} = 1|X_i, D_i = 1) - E[Y_i(0)|X_i = x] \\
&= \sum_{t \neq 0} E[Y_i(t)|D_{t,i} = 1, X_i = x] \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} - E[Y_i(0)|X_i] \\
&= \sum_{t \neq 0} \underbrace{E[Y_i(t) - Y_i(0)|X_i = x]}_{t-\text{specific CATE}} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \\
&\quad + \sum_{t \neq 0} \underbrace{\{E[Y_i(t)|D_{t,i} = 1, X_i = x] - E[Y_i(t)|X_i = x]\}}_{\text{selection effect}} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)}
\end{aligned} \tag{3}
$$

Equation (3) shows that the estimand consists of two components: First, a weighted average of $CATEs$ of the effective treatments, $\tau_t(x) = E[Y_i(t) - Y_i(0)|X_i = x]$, with weights depending on the conditional probability of being treated in the respective effective treatment. Second, a weighted average of effective treatment specific selection effects. The selection effects are positive if those with characteristics x who are actually observed in treatment t show higher potential outcomes than the general population described by x, or negative if vice versa. This term is relevant if there is selection into the effective treatment even after conditioning on the observed confounders. This can e.g. occur in the case of a

randomized binary treatment in Scenario 2 where the selected X_i might not include all confounders for the treatment versions.

The decomposition in (3) highlights that the interpretation of the underlying estimand becomes more nuanced in the presence of heterogeneous treatments. What is supposed to be an easily interpretable $CATE$ depends now on the potentially unknown distribution of effective treatments and selection into those treatments. Thus, without further assumptions, heterogeneous effects attributed to the binary indicator can be driven by different $CATEs$, different compositions of the effective treatments, different selection effects of the effective treatments, or combinations thereof.

To be able to meaningfully decompose estimand (3) below, we impose a strong ignorability assumption at the effective treatment level:

Assumption 1 *(strong ignorability of effective treatment)*

(a) Unconfoundedness: $Y_i(t) \perp\!\!\!\perp D_{t,i}|X_i = x, \forall\ t \in \mathcal{T}$ *and* $x \in \mathcal{X}$.

(b) Common support: $0 < P[D_{t,i} = 1|X_i = x] \equiv e_t(x), \forall\ t \in \mathcal{T}$ *and* $x \in \mathcal{X}$.

Assumption 1 is a standard assumption in the multiple treatments setting (Imbens, 2000; Lechner, 2001). It imposes that (a) the set of conditioning variables is rich enough such that after conditioning all residual variation in potential outcomes is independent of the allocated effective treatment and (b) there are comparable units across effective treatments in terms of their confounders. Under this assumption, the selection effects in (3) disappear and the underlying estimand becomes

$$\tau(x) = \sum_{t \neq 0} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \tau_t(x) \equiv nATE(x). \tag{4}$$

We call this estimand the *natural conditional average treatment effect* $nATE(x)$ because it is the result of the actual or "natural" effective treatment composition. It is important to note that, even under Assumption 1, the differences between units characterized by x and x' can result from different treatment shares, different treatment $CATEs$, or both. We thus could detect seemingly heterogeneous effects, even if the treatment $CATEs$ are constant within treatments but not homogeneous between treatments, i.e. $\tau_t(x) = \tau_t\ \forall\ t \in \mathcal{T}, x \in \mathcal{X}$ but $\tau_t \neq const.\ \forall\ t \in \mathcal{T}$, as long as the probabilities to be observed in the different

effective treatments are heterogeneous. In this case any difference is driven by treatment heterogeneity:

$$nATE(x) - nATE(x') = \sum_{t \neq 0} \left[\frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} - \frac{e_t(x')}{\sum_{t \neq 0} e_t(x')} \right] \tau_t \tag{5}$$

This fundamentally affects the interpretation of heterogeneous effects even if the underlying effective treatments are not observable. If they are observable, however, we can further decompose heterogeneous effects of the binary indicator in what follows.

3.3 The Decomposition

In this section we demonstrate how to disentangle actual effect heterogeneity and heterogeneity driven by selection into effective treatments. We propose to decompose the $nATE(x)$ in two parts:

$$\underbrace{\sum_{t \neq 0} \frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} \tau_t(x)}_{nATE(x)} = \underbrace{\sum_{t \neq 0} \frac{\pi_t}{\sum_{t \neq 0} \pi_t} \tau_t(x)}_{rATE(x)} + \underbrace{\sum_{t \neq 0} \left(\frac{e_t(x)}{\sum_{t \neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t \neq 0} \pi_t} \right) \tau_t(x)}_{\Delta(x)} \tag{6}$$

where $\pi_t = E[D_{t,i}]$ are the unconditional treatment probabilities. The first component on the right hand side fixes the composition of the effective treatments at the population value. It resembles a situation where effective treatments are randomly allocated using the population level selection probabilities. Thus, we refer to it as the *random conditional average treatment effect* $rATE(x)$. All heterogeneity in $rATE(x)$ is driven by "real" effect heterogeneity within treatments, $\tau_t(x) \neq \tau_t(x')$ for some $x, x' \in \mathcal{X}$, as the underlying treatment composition is held fixed. In other words, differences in $rATE(x)$ describe effect heterogeneity *compositionis paribus*. Thus, we can exploit potential heterogeneity in $rATE(x)$ to test necessary conditions for classic (or "within") effect heterogeneity.

The second component of the decomposition $\Delta(x)$ is the part of $nATE(x)$ stemming from the interaction of non-constant effective treatment probabilities and different effective treatments having different effects ("between" treatment effect heterogeneity). Thus, the decomposition is redundant, i.e. $\Delta(x) = 0 \; \forall \; x \in \mathcal{X}$, under (i) effective treatment

composition homogeneity $\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0}\pi_t} = 0 \; \forall \, t \in \mathcal{T}$ and $x \in \mathcal{X}$, (ii) treatment variation irrelevance $E[Y_i(t)|X_i = x] = E[Y_i(t')|X_i = x] \; \forall \; x \in \mathcal{X}, \; t, t' \in \mathcal{T}$ of VanderWeele (2009)[4], or (iii) if positive and negative components net out to zero. Hence, $\Delta(x) \neq 0$ is a necessary condition for unequal treatment probabilities and between treatment effect heterogeneity. Furthermore, heterogeneity in $\Delta(x)$ is a necessary condition for heterogeneous assignment probabilities, within treatment effect heterogeneity, or both. Thus, the decomposition can be used to address a variety of relevant policy questions, see also Remark 2 below. Moreover, the focus on such necessary conditions offers statistical advantages over testing similar conditions in the standard multi-valued treatment effect setup when there are *many* effective treatments. We return to this point in Remark 3 below and in Section 5.

Under Assumption 1, the conditional average potential outcome of treatment t is identified as $\mu_t(X_i) \equiv E[Y_i(t)|X_i] = E[Y_i(t)|D_{t,i} = 1, X_i] = E[Y_i|D_{t,i} = 1, X_i]$ and accordingly the decomposition terms are identified as:

$$
\begin{aligned}
nATE(x) &= \sum_{t\neq 0} \frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} (\mu_t(x) - \mu_0(x)) \\
rATE(x) &= \sum_{t\neq 0} \frac{\pi_t}{\sum_{t\neq 0} \pi_t} (\mu_t(x) - \mu_0(x)) \\
\Delta(x) &= \sum_{t\neq 0} \left(\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0} \pi_t} \right) (\mu_t(x) - \mu_0(x))
\end{aligned} \tag{7}
$$

Aggregations or projections of the three estimands are thus also identified. In particular, let Z_i denote a (low dimensional) subset of confounders supported on $\mathcal{Z} \subset \mathcal{X}$ and define

$$
\begin{aligned}
nATE(z) &= E[nATE(X_i)|Z_i = z] \\
rATE(z) &= E[rATE(X_i)|Z_i = z] \\
\Delta(z) &= E[\Delta(X_i)|Z_i = z].
\end{aligned} \tag{8}
$$

These parameters provide more concise, predictive summaries of heterogeneity or allocation differences for specific subgroups defined by $Z_i = z$. The unconditional decomposition

[4]In this case $\tau_t(x) = \tau(x)$ and consequently $\Delta(x) = \tau(x) \underbrace{\sum_{t\neq 0} \left(\frac{e_t(x)}{\sum_{t\neq 0} e_t(x)} - \frac{\pi_t}{\sum_{t\neq 0}\pi_t} \right)}_{=0} = 0, \; \forall \; x \in \mathcal{X}$.

terms $nATE = E[nATE(X_i)]$, $rATE = E[rATE(X_i)]$, and $\Delta = E[\Delta(X_i)]$ are special cases thereof. We propose an estimation and inference method for these parameters in Section 4.

Remark 1: In principle, an analogous decomposition could also be constructed with alternative weights for the effective treatments, e.g. equal weighting $1/J$. However, using the unconditional effective treatment probabilities ensures that $nATE(x) = rATE(x)$ in the case of completely randomized effective treatments.

Remark 2: The interpretation of $\Delta(x)$ depends on the scenario:

- *Scenario 1:* $\Delta(x)$ and its aggregates have a descriptive interpretation. It describes how much of $nATE(x)$ is driven by an underlying effective treatment mix that deviates from the population mix. A non-constant $\Delta(x)$ indicates that the binarization has consequences for the detected heterogeneous effects. Thus, it helps to understand heterogeneity resulting from the binarization.

- *Scenario 2:* $\Delta(x)$ and its aggregates provide information for assignment evaluation. Positive values indicate that the assignment of treatment versions is better than random. Negative values indicate worse than random version assignment assuming that individuals act equivalently under the hypothetical random assignment compared to the observational assignment (Heckman, 2020). A non-constant $\Delta(x)$ indicates that the selection quality of versions varies across different groups. Thus, the estimand provides an evaluation of the actual assignment mechanism.

Remark 3: In principle, one could be tempted to test effect heterogeneity within the classic multi-valued treatment framework by comparing up to $J(J-1)/2$ conditional average treatment effects. This requires correction for multiple testing. More importantly, however, the required strong overlap assumption for valid regular inference is increasingly difficult to justify when J is large. Our decomposition approach, on the other hand, is robust to too many treatments and limited overlap. In particular, even when $J \to \infty$, strong overlap assumptions for accurate inference based on asymptotic normality are not required. This comes at the cost of testing somewhat weaker conditions regarding effect heterogeneity: In the multi-valued treatment effect setting we can test *sufficient* conditions

for effect heterogeneity between and within all treatments, while the decomposition allows for testing *necessary* conditions for within and between treatment effect heterogeneity.

Remark 4: The comparison of $nATE(x)$ and $rATE(x)$ shows resemblance to the relationship between the ATE and the average treatment effect on the treated ($ATET$) in the canonical setting with a homogeneous binary treatment. The $ATET$ gives the average effect of those treated under the actual treatment assignment, while the ATE gives the average effect under the hypothetical random assignment of treatment.

Remark 5: The unconditional $rATE$ is a special case of the composite treatment effects described in Lechner (2002). Furthermore, if we allow for $J \to \infty$, it can approximate the integrated dose-response function of a continuous treatment as defined in Kennedy, Ma, McHugh, and Small (2017), see Section B.3 for more details. The unconditional Δ is also similar in spirit to the Population Average Prescriptive Effect defined by Imai and Li (2021) in the context of policy learning.

4 Estimation and Inference

In this section, we outline a flexible estimation approach for the (conditional) decomposition terms and propose a method for conducting valid statistical inference. The method accommodates the use of modern machine learning and other non- or semiparametric methods in the estimation of the required nuisance parameters.

We propose to approximate the conditional expectations of the decomposition terms $g(z)$ by a linear combination of transformations $b(z)$ of heterogeneity variables z, i.e.

$$g(z) = b(z)'\beta_0 + r_g(z) \tag{9}$$

where β_0 is the parameter vector of the best linear predictor given as solution to equation $E[b(Z_i)(g(Z_i) - b(Z_i)'\beta_0)] = 0$. $r_g(z)$ is the approximation error and $b(z)$ can be basis transformations of the regressors of interest such as polynomials, splines, wavelets, or other functions. The number of components in $b(\cdot)$ is allowed to grow with the sample size which allows us to be agnostic about the shape of the true g-function.

Let in the following $\eta = \eta(x) = (\mu_0(x), \ldots, \mu_J(x), e_0(x), \ldots, e_J(x))'$ denote the vector

of nuisance quantities and write $\eta = \eta_i = \eta(X_i)$ with subscript and argument suppressed whenever it does not cause confusion. Also define the unconditional selection probability vector $\pi = (\pi_0, \ldots, \pi_J)$.

Table 1: Score Functions of the Decomposition Parameters

Parameter	Score function $\psi_i(\eta, \pi) = \psi_i^{[Parameter]}(\eta, \pi)$
$nATE$	$\Psi_i(\eta) - \psi_i^{[0]}(\eta)$
$rATE$	$\frac{\sum_{t\neq 0} \pi_t \psi_i^{[t]}(\eta)}{\sum_{t\neq 0} \pi_t} - \psi_i^{[0]}(\eta)$
Δ	$\Psi_i(\eta) - \frac{\sum_{t\neq 0} \pi_t \psi_i^{[t]}(\eta)}{\sum_{t\neq 0} \pi_t}$

The scores $\psi_i^{[t]}(\eta)$ and $\Psi_i(\eta)$ are defined in equations (10) and (11), respectively.

We follow the general idea of Semenova and Chernozhukov (2021) to construct "Neyman-orthogonal" scores $\psi_i(\eta, \pi)$ such that $g(z) = E[\psi_i(\eta, \pi)|Z_i = z]$. These scores are defined by having an (approximate) zero Gateaux derivative with respect to the underlying nuisance parameters at the true parameter vector (Chernozhukov et al., 2018). The robust scores for the three decomposition parameters considered here are weighted combinations of the well-known Neyman-orthogonal scores for average potential outcomes (Robins & Rotnitzky, 1995), also known as augmented inverse probability weighting (AIPW) scores:

$$\psi_i^{[t]}(\eta) = \mu_t(X_i) + \frac{D_{t,i}(Y_i - \mu_t(X_i))}{e_t(X_i)} \tag{10}$$

$$\begin{aligned}\Psi_i(\eta) &= E[Y_i|D_i = 1, X_i] + \frac{D_i(Y_i - E[Y_i|D_i = 1, X_i])}{P(D_i = 1|X_i)} \\ &= \frac{\sum_{t\neq 0} \mu_t(X_i) e_t(X_i)}{\sum_{t\neq 0} e_t(X_i)} + \frac{D_i\left[Y_i - \frac{\sum_{t\neq 0} \mu_t(X_i) e_t(X_i)}{\sum_{t\neq 0} e_t(X_i)}\right]}{\sum_{t\neq 0} e_t(X_i)}\end{aligned} \tag{11}$$

where $\psi_i^{[t]}(\eta)$ is the score of the treatment t specific average potential outcome and $\Psi_i(\eta)$ is the score for the group described by the binary indicator. Table 1 shows how to combine these scores to form unbiased signals of the decomposition parameters. These combinations retain Neyman-orthogonality with respect to η, see Appendix B.4, but inference has to be adjusted for uncertainty in the estimation of π, see Section 5.

Consider now the regression of the score functions onto the space spanned by the k-dimensional transformation of Z_i, $b(Z_i)$. This yields the estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n} b(Z_i) b(Z_i)' \right)^{-1} \sum_{i=1}^{n} b(Z_i) \psi_i(\hat{\eta}, \hat{\pi}) \tag{12}$$

where the score of a decomposition term with estimated nuisance quantities $\psi_i(\hat{\eta}, \hat{\pi})$ serves as pseudo-outcome in the corresponding least squares regression on $b(Z_i)$. For $\hat{\pi}$ we use simple sample averages, i.e. $\hat{\pi}_t = n^{-1} \sum_{i=1}^{n} D_{t,i}$. Estimation of $\hat{\eta}$ can be done via modern machine learning methods such as random forests, deep neural networks, high-dimensional sparse likelihood and regression models or other non- and semiparametric estimation methods with good approximation qualities for the functions at hand. For details regarding the technical assumptions, consider Section 5. We require that all components in $\hat{\eta}$ are obtained via K-fold cross-fitting:

Definition 4.1 ***K-fold cross-fitting*** *(see Definition 3.1 in Chernozhukov et al. (2018))* *Take a K-fold random partition $(I_f)_{f=1}^{K}$ of observation indices $[K] = \{1, \ldots, n\}$ with each fold size $n_f = n/K$. For each $f \in [K] = \{1, \ldots, K\}$, define $I_f^c := \{1, \ldots, n\} \backslash I_f$. Then for each $f \in [K]$, the machine learning estimator of the nuisance function are given by*

$$\hat{\eta}_f = \hat{\eta}((Y_i, X_i, T_i)_{i \in I_f^c}).$$

Thus for any observation $i \in I_f$ the estimated score only uses the model for η learned from the complementary folds $\psi_i(\hat{\eta}, \hat{\pi}) = \psi_i(\hat{\eta}_f, \hat{\pi})$.

Cross-fitting controls the potential bias arising from overfitting using flexible machine learning methods without the need to evaluate entropy conditions for the function class that contains true and estimated nuisance quantities. If finite parametric models such as linear or logistic are assumed for the nuisance quantities, the proposed methodology can be applied without the need for cross-fitting.

Under suitable assumptions, the predictions using estimator (12) are consistent for $g(z)$. Moreover, it is possible to conduct asymptotically valid inference around the best linear predictor, i.e. for any $z_0 = z_{0,n}$ we can construct $(1 - \alpha)\%$ confidence intervals for

the true decomposition function as

$$CI_{1-\alpha}(g(z_0)) = \left[b(z_0)'\hat{\beta} \pm q_{1-\alpha/2} \sqrt{\frac{b(z_0)'\hat{\Omega} b(z_0)}{n}} \right] \tag{13}$$

where $q_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution and $\hat{\Omega}$ is a consistent sample estimator of the asymptotic variance Ω (see Section 5 and Appendix B.5). The estimator explicitly takes into account the additional uncertainty from estimating the unconditional treatment probabilities in the decomposition terms. The interval in (13) is also valid for the best linear predictor $b(z_0)'\beta_0$ under moderate misspecification if the approximation error is not too large. It provides asymptotically accurate confidence intervals around the true g-function if the approximation error vanishes at a suitable rate as the number of basis functions or transformations increases. For the technical details consider Section 5.

5 Large Sample Properties

5.1 Assumptions and Main Results

In this section, we present and discuss the large sample properties of the proposed decomposition method. First, we introduce the relevant definitions. We then discuss the assumptions required for (i) all decomposition parameters, (ii) $nATE$, and (iii) $rATE/\Delta$ specifically and their connections to the literature. We contrast (ii) and (iii) as the $nATE$ tends to require less restrictive conditions compared to $rATE/\Delta$. We then present the main Theorem and outline potential extensions.

In the following, quantities like $\psi_i()$ or $r_g()$ are used in their generic sense, i.e. for a given choice of decomposition parameter $nATE$, $rATE$ or Δ. $a \lesssim b$ means $a/b = O(1)$ and $a \lesssim_P b$ means $a/b = O_p(1)$. For a general matrix M denote its largest (smallest) eigenvalue by $\lambda_{max}(M)$ ($\lambda_{min}(M)$). The realization sets of the estimated nuisance quantities are given by $\mathcal{H}_n = \mathcal{E}_n \times \mathcal{M}_n$ with $\mathcal{E}_n = E_{0,n} \times E_{1,n} \times \cdots \times E_{J,n}$ and $\mathcal{M}_n = M_{0,n} \times M_{1,n} \times \cdots \times M_{J,n}$, where $E_{t,n}$ and $M_{t,n}$ are the realization sets for e_t and μ_t respectively. For estimators

$\hat{e}_t(X_i)$ and $\hat{\mu}_t(X_i)$ define the L_q error rates

$$s_{t,n,q} = \sup_{\hat{e}_t \in E_{t,n}} E[(\hat{e}_t(X_i) - e_t(X_i))^q]^{1/q}, \quad m_{t,n,q} = \sup_{\hat{\mu}_t \in M_{t,n}} E[(\hat{\mu}_t(X_i) - \mu_t(X_i))^q]^{1/q}$$

and the slowest L_q rates over all treatments $s_{n,q} = \sup_{t \neq 0} s_{t,n,q}$ and $m_{n,q} = \sup_{t \neq 0} m_{t,n,q}$. Note that $\psi_i^{[t,0]}(\eta) = \psi_i^{[t]}(\eta) - \psi_i^{[0]}(\eta)$. By definition $\varepsilon_i = \psi_i(\eta, \pi) - E[\psi_i(\eta, \pi)|Z_i]$ where $\psi_i(\eta, \pi)$ corresponds to the moment function of the decomposition parameter of choice from Table 1. Denote $g(z) = E[\psi_i(\eta, \pi)|Z_i = z]$ where $g \in \mathcal{G}$ with $\mathcal{G} = \mathcal{G}_n$ being a function class potentially depending on n. Thus, $g(z) = b(z)'\beta_0 + r_g(z)$ where β_0 is the parameter of the best linear predictor defined as the root of equation $E[b_i(g(Z_i) - b_i'\beta_0)] = 0$ where $b_i = b(Z_i)$. Also define the potential outcome mean error $\varepsilon_i(t) = Y_i(t) - E[Y_i(t)|X_i]$ and its conditional variance $\sigma_t^2(X_i) = E[\varepsilon_i(t)^2|X_i]$. For the $nATE$ machine learning bias components, we define

$$B_n^{[nATE]} := \sqrt{n} \sup_{\hat{\eta} \in H_n} ||E[b_i(\psi_i^{[nATE]}(\hat{\eta}, \pi) - \psi_i^{[nATE]}(\eta, \pi))]||$$
$$\Lambda_n^{[nATE]} := \sup_{\hat{\eta} \in H_n} (E[||b_i(\psi_i^{[nATE]}(\hat{\eta}, \pi) - \psi_i^{[nATE]}(\eta, \pi))||^2])^{1/2}$$

and equivalently for $rATE/\Delta$ with moment functions according to Table 1. Remainder terms R_n are defined in Appendix A. Let $\gamma_t = E[b_i\psi_i^{[t,0]}(\eta)] = E[b_i\tau_t(X_i)]$, $\gamma = (\gamma_1 \dots \gamma_J)$, and define $a_i = (a_i^{[1]} \dots a_i^{[J]})'$ with $a_i^{[t]} = (1 - \pi_0)^{-2}(D_{t,i}(1 - \pi_0) + D_{0,i}\pi_t - \pi_t)$. Now let $Q = E[b_i b_i']$ and define

$$\Omega = Q^{-1}E[(b_i(\varepsilon_i + r_i) + \gamma a_i)(b_i(\varepsilon_i + r_i) + \gamma a_i)']Q^{-1}$$
$$\Omega_1 = Q^{-1}E[b_i b_i'(\varepsilon_i + r_i)^2]Q^{-1}$$
$$\Omega_2 = Q^{-1}E[\gamma a_i a_i'\gamma']Q^{-1}.$$

We now present the assumptions required for all decomposition parameters. They are meant to hold uniformly over n if not stated otherwise:

Decomposition Assumptions:

A.1) (Identification) Q has eigenvalues bounded above and away from zero.

A.2) (Conditional means) The potential outcomes have bounded conditional means

$$\sup_t \sup_{x \in \mathcal{X}} \mu_t(x) \lesssim 1$$

A.3) (Control overlap and limited treatment overlap) The control propensities are bounded away from zero, i.e. for some $c \in (0, 1/2)$

$$c < \inf_{x \in \mathcal{X}} e_0(x) \leq \sup_{x \in \mathcal{X}} e_0(x) < 1 - c$$

and the re-scaled inverse treatment propensity scores are proportional to the number of different treatments

$$\sup_{t \neq 0} \sup_{x \in \mathcal{X}} \frac{\pi_t}{e_t(x)} \lesssim 1 \quad \textit{and} \quad J\pi_t \lesssim 1$$

for all $t = 1, \ldots, J$ with $J = o(n)$.

A.4) (Bounded relative prediction error) On the realization set with probability $1 - u_n$, the worst relative prediction error for the cross-fitted treatment propensities are bounded

$$\sup_{t \neq 0} \sup_{\hat{e}_t \in E_{t,n}} \sup_{x \in \mathcal{X}} \frac{e_t(x)}{\hat{e}_t(x)} \lesssim 1$$

A.1 rules out multicollinearity of the basis functions used for the nonparametric heterogeneity analysis in the last stage. A.2 is a mild heterogeneity restriction on the potential outcomes. A.3 is crucial: It is concerned with the degree of overlap for a general number of treatments $J = J_n$. In particular, it assumes that there is strong overlap for the control group and the aggregate treatment, i.e. control and aggregate treatment propensities are uniformly bounded away from zero. However, the propensities for treatments $t = 1, \ldots, J$ are allowed to be arbitrarily close to zero as long as they vanish at most at a rate proportional to their respective unconditional treatment selection probability π_t. This allows for limited overlap at each treatment level which is necessary when their number is allowed to increase with the sample size, i.e. $J \to \infty$. We suggest to assess Assumption A.3 empirically by analyzing the (estimated) distribution of $e_t(x)/\pi_t$ for all t: If these re-scaled scores have sufficient density bounded away from zero by the same standard used to assess conventional propensity score distributions (Heiler & Kazak, 2021), then the assumption is likely to hold, see Appendix B.8 and B.9 for examples based on the empirical applications from Section 7. Assuming a homogeneous $1/J$-rate for all π_t is without loss of generality: If the product converges to zero for some t, it vanishes from relevant first-order approximations and estimation properties are eventually

determined by the treatments that obey Assumption A.3. Moreover, if A.3 only applies to a smaller finite subset of treatments, it effectively corresponds to strong overlap for these particular t and thus estimators behave analogously to standard AIPW for a control potential outcome. Note that the growth of J is restricted to rate $o(n)$ such that consistent estimation of unconditional multi-valued treatment effects is still possible, albeit at a slower rate compared to the strong overlap case similar to Hong et al. (2020). If there are many control conditions aggregated into $D_i = 0$, then strong overlap for controls could also be relaxed analogously to the re-scaled treatment propensities.

A.4 says that the worst relative prediction for the cross-fitted propensities is bounded on the realization set. This is a non-standard assumption, in particular when $J \to \infty$. It is likely to hold for frequency based methods, i.e. estimators that use some form of (weighted) average within the cells defined by $D_{t,i}$ for $t = 0, \ldots, J$ to construct propensities including advanced machine learning methods. A sufficient, but by no means necessary, condition is uniform consistency of $\hat{e}(x)$ over $\mathcal{X}$ at rate $o(e_t(x)^{-1})$. This can be shown to hold for single-index models (Ma, Sasaki, & Wang, 2022) and nonparametric kernel regression (Heiler & Taylor, 2022) under weak conditions. The key point is that these estimators inherit a *local superefficiency* property from $\hat{\pi}_t$, i.e. faster convergence rate $|\hat{\pi}_t - \pi_t| \lesssim_P (nJ)^{-1}$ in regimes with many treatments/vanishing unconditional selection probabilities. A.4 then requires the estimators to have a consistency rate increased by a factor of $\sqrt{J}$ compared to the finite J case. For parametric estimators this holds as long as $J = o(n)$ while for kernel regression, for example, under the usual smoothness assumptions with $\mathcal{X}$ of dimension d and a bandwidth h, it requires that $\sqrt{\log(n) n h^d / J} = o(1)$ (Heiler & Taylor, 2022). We provide some more intuition about Assumption A.4 and the links between large J and superefficient nuisance parameter estimation in Section 5.2.

We now present the assumptions required for $nATE$ followed by $rATE/\Delta$:

$nATE$ **Assumptions:**

For some $m > 2$, we have that:

B.1) (Conditional Moments) The potential outcomes have at least m conditional moments for the treated: $\sup_t \sup_{z \in \mathcal{Z}} E[\varepsilon_i(t)^m | Z_i = z, D_{t,i} = 1] \lesssim 1$.

B.2) (Approximation) For each n and k, there are finite constants c_k and l_k such that for

each $g \in \mathcal{G}$

$$||r_g||_{P,2} := \sqrt{\int_{z\in\mathcal{Z}} r_g^2(z)dP(z)} \leq c_k,$$
$$||r_g||_{P,\infty} := \sup_{z\in\mathcal{Z}} |r_g(z)| \leq l_k c_k.$$

B.3) (Machine Learning Bias) For some $h_1, h_2 > 0$ *with* $1/h_1 + 1/h_2 = 1$ *we have that*

$$B_n^{[nATE]} \lesssim \sqrt{nk} s_{0,n,2h_2}\left(J s_{n,2h_1} + m_{n,2h_1}\right) = o(1),$$
$$\Lambda_n^{[nATE]} \lesssim \xi_k\left(s_{0,n,2} + \sqrt{J} s_{n,2} + J^{-1} m_{n,2}\right) = o(1).$$

B.4) (Basis and Linearization Error) The k *basis functions are chosen such that*

$$||R_{n,Q}|| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}}\left(1 + k^{1/2} l_k c_k\right) = o(1).$$

B.5) (Basis and Lindeberg Condition) Let $\sqrt{n}/\xi_k - l_k c_k \to \infty$ *such that*

$$\frac{J}{[\sqrt{n}/\xi_k - l_k c_k]^2} + \left(\frac{(l_k c_k)^{\frac{2}{m}} J^{\frac{1}{m}}}{[\sqrt{n}/\xi_k - l_k c_k]}\right)^m = o(1).$$

B.1 imposes some regularity on the tails of the conditional potential outcomes. B.2 defines the L_2 and uniform approximation rates using the basis functions for function class $\mathcal{G}$. If the basis is sufficiently rich to span $\mathcal{G}$, we say it is correctly specified and $c_k \to 0$ as $k \to \infty$. However, our results allow for the case of misspecification, i.e. $c_k \not\to 0$. This is a standard characterization in the literature on nonparametric series methods, see e.g. Belloni et al. (2015) for more details and examples.

B.3 is crucial: It requires high-quality approximation capabilities of the first-stage machine learning methods for the nuisance quantities. In the case of a finite-dimensional, bounded basis $\sup_{z\in\mathcal{Z}} ||b(z)||_\infty < C$ and finite J, the conditions can be simplified to $\sqrt{nk} s_{0,n,2}(s_{n,2} + m_{n,2}) = o(1)$. This means that the products of the nuisance quantities for the conditional control propensity and treatment propensities/potential outcome means have to converge at least at rate $o((nk)^{-1/2})$ identical to conditional ATE estimation in Semenova and Chernozhukov (2021). This flexible rate requirement is a consequence of the Neyman-orthogonality of the moment function with regards to the nuisance parameters η. In the simple case of a fixed basis (e.g. a constant univariate basis as in unconditional

binary ATE estimation) it reduces to the well-known requirement in the double/de-biased machine learning literature that the nuisance functions have to be of rate $o(n^{-1/4})$ (Chernozhukov et al., 2018). For many treatments $J \to \infty$, flexible k, and/or machine learning estimators, the convergence requirements can be more demanding. We discuss these cases and corresponding rate requirements in Section 5.2.

B.4 controls the approximation error from linearization of the estimator taking into account the unknown design matrix Q of increasing dimension. The condition is equivalent to the one required for linearization in conventional least squares series estimation (Belloni et al., 2015). This suggests that, for more specific series methods such splines (Huang, 2003) and local partitioning estimators (Cattaneo et al., 2020), the rate can be improved to $\sqrt{\xi_k^2 \log k/n}(1 + \sqrt{\log k} l_k c_k)$, see also Belloni et al. (2015), Section 4 and Cattaneo et al. (2020), Remark SA-4 (page 12) of their supplemental appendix. Note that this rate does not depend on J as, for linearization, the treatment dimension enters only through estimation of the expanding set of nuisance parameters. Once the difference between true and estimated nuisance parameters is controlled for via B.3, there is no difference to the standard series estimation/binary ATE case with no or known nuisances.

B.5 controls the rate of the basis function relative to approximation error such that the Lindeberg condition for asymptotic normality holds. Note that this rate is required to be faster by a factor of J relative to conventional series estimation. This is due to the fact, that the tails of the summands that determine the first-order asymptotic distribution are selected from a combination of J different potential outcome errors $\varepsilon_i(t)$ for $t = 1, \ldots, J$. Thus, the conditions for the many treatments case are somewhat stronger then the ones expected for series estimation or (conditional) ATE estimation under a moment assumption such as B.1 and finite J.

$rATE/\Delta$ **Assumptions**

For some $m > 2$, we have that:

C.1) (Conditional Moments) The potential outcomes have at least m conditional moments for the selected: $\sup_t \sup_{z \in \mathcal{Z}} E[\varepsilon_i(t)^m | Z_i = z, D_{t,i} = 1] \lesssim 1$.

C.2) (Approximation) For each n and k, there are finite constants c_k and l_k such that for

each $g \in \mathcal{G}$

$$||r_g||_{P,2} := \sqrt{\int_{z\in\mathcal{Z}} r_g^2(z)dP(z)} \leq c_k,$$
$$||r_g||_{P,\infty} := \sup_{z\in\mathcal{Z}} |r_g(z)| \leq l_k c_k.$$

C.3) (Machine Learning Bias) For some $h_1, h_2 > 0$ *with* $1/h_1 + 1/h_2 = 1$ *we have that*

$$\begin{aligned} B_n^{[rATE]} &\lesssim \sqrt{nk} J s_{n,2h_1} m_{n,2h_2} && = o(1), \\ \Lambda_n^{[rATE]} &\lesssim \xi_k(m_{n,2} + J s_{n,2} + \sqrt{J s_{n,2} m_{n,2h_1} m_{n,2h_2}}) && = o(1). \end{aligned}$$

C.4) (Basis and Linearization Error) The k *basis functions are chosen such that*

$$\begin{aligned} ||R_{n,\pi}|| &\lesssim_P J\sqrt{k}(n^{-1/2} + m_{n,2} + J s_{n,2}) && = o(1), \\ ||R_{n,Q}|| &\lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}}\left(1 + k^{1/2}(J^{1/4} + l_k c_k)\right) && = o(1). \end{aligned}$$

C.5) (Basis and Lindeberg Condition) Let $n/kJ^2 \to \infty$, $\sqrt{n}/\xi_k - l_k c_k \to \infty$ *such that*

$$\frac{J^4}{[\sqrt{n}/\xi_k - l_k c_k]^2} + \left(\frac{(l_k c_k)^{\frac{2}{m}} J}{[\sqrt{n}/\xi_k - l_k c_k]}\right)^m = o(1).$$

C.6) (Eigenvalues) $\lambda_{min}(\Omega_2) > 0$ *and* $\lambda_{max}(\Omega_1)/\lambda_{min}(\Omega) \lesssim 1$.

We discuss and contrast Assumptions C.1–C.6 with B.1–B.5: C.1 and C.2 are equivalent to B.1 and B.2 with potentially different m, c_k, and l_k. C.3 controls for the estimation of nuisance parameters. In the case of a bounded basis, the condition for B_n reduces to $\sqrt{nk} J s_{n,2} m_{n,2} = o(1)$ which is expected to be equivalent to the $nATE$ machine learning bias rate when J is finite. However, in the large J case, estimating the potential outcome means at rate $m_{n,2}$ is generally slower than estimating the control propensities at $s_{0,n,2}$. In the parametric case, for example, we have that $m_{n,2} = \sqrt{J/n} = \sqrt{J} s_{0,n,2}$, see Section 5.2. An equivalent argument holds for Λ_n also leading to an additional $\sqrt{J}$ factor compared to the $nATE$ case (here the third term in Λ_n is dominated by the first two and can be ignored). Thus, the product rates have to be faster by a factor of $\sqrt{J}$ in this case, i.e. $rATE/\Delta$ generally require somewhat higher quality first-stage learners in comparison to the $nATE$.

C.4 provides the error from the linearization. First note that there is an additional term $R_{n,\pi}$ due to the moment functions not being Neyman-orthogonal with respect to the

unconditional weights π used for $rATE/\Delta$. It puts an additional restriction on the growth of the number of treatments. For example, the first condition reduces to $\sqrt{kJ^3/n} = o(1)$ in case of parametric nuisance quantities. $R_{n,Q}$ corresponds to the $nATE$ case plus an additional term of order $\sqrt{\xi_k \log k/n} k^{1/2} J^{1/4}$. This is a result of the interaction between estimation error from estimating the unconditional weights with design matrix Q. Again, for specific series such as splines or local partitioning, we conjecture that a faster rate of $\sqrt{\xi_k \log k/n}\sqrt{\log k} J^{1/4}$ is attainable. Note that the additional factors are only of order $J^{1/2}$ and $J^{1/4}$ compared to the $nATE$. This is due to the superefficiency of the unconditional probability estimates $\hat{\pi}_t$ in the $J \to \infty$ case.

C.5 is similar to B.5 but requires more stringent conditions on J and ξ_k compared to the $nATE$. The Lindeberg condition for $nATE/\Delta$ is driven by the tails of a weighted combination of moment functions from many treatment groups which can have high variance when J is large. It is more restrictive compared to the $nATE$, as, for the latter, the weight for each t-specific moment function $\psi_i^{[t,0]}(\eta)$ is the actual treatment propensity $e_t(X_i)$, see Table 1. Thus, the inverse propensity score weights disappear leading to a lower variance for the $nATE$[5] explaining the additional J-dependent factors between B.5 and C.5.

The first condition in C.6 rules out the degenerate case where $nATE = rATE$. Naturally, if this is true, Assumptions B.1–B.5 apply instead. The second condition excludes the hypothetical case where the sum of noise plus approximation error is perfectly negatively correlated with the (γ-weighted) error from estimating the unconditional weights π. Both restrictions are expected to always hold in practice and can also be assessed by looking at the empirical analogues of Ω, Ω_1, and Ω_2.

For the estimation of the asymptotic variance, we also assume that A.V holds. The corresponding details and discussion can be found in Appendix B.6.

A.V) (Asymptotic Variance) The assumptions in Appendix B.6 hold for nATE and rATE or Δ respectively, i.e. $||\hat{\Omega} - \Omega|| = o_p(1)$.

[5]This is somewhat related to Li, Morgan, and Zaslavsky (2018) who show that propensity score weighted average treatment effects can be estimated with smallest variance within a class of effect parameters. In their case the weights are $e(X_i)(1 - e(X_i))$ and they only consider the binary treatment case, but a similar intuition applies in the framework of this paper.

A.V can require somewhat stronger moment and growth conditions for the basis and/or number of treatments. For example, for the $nATE$, they reduce to the same rates required by Semenova and Chernozhukov (2021), Theorem 3.3, condition (ii) with factor $n^{1/m}$ replaced by $(nJ)^{1/m}$. Under finite J, they are again equivalent. We obtain the following Theorem:

Theorem 5.1 *Let $\Phi(\cdot)$ denote the Gaussian cumulative distribution function. Suppose Assumptions A.1 – A.4, A.V, and B.1 – B.5 (C.1 – C.6) hold for nATE (rATE/Δ) and $\hat{\beta}$ and $\hat{\Omega}$ are estimated according to (12) and (18) respectively. Then, for any $z_0 = z_{0,n}$,*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n} \frac{b(z_0)'(\hat{\beta} - \beta_0)}{\sqrt{b(z_0)'\hat{\Omega} b(z_0)}} \leq t \right) - \Phi(t) \right| = 0.$$

Moreover if the approximation error is small, $\sqrt{n} r_g(z_0) / \sqrt{b(z_0)'\Omega b(z_0)} \to 0$, then

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| P\left(\sqrt{n} \frac{b(z_0)'\hat{\beta} - g(z_0)}{\sqrt{b(z_0)'\hat{\Omega} b(z_0)}} \leq t \right) - \Phi(t) \right| = 0.$$

Theorem 5.1 demonstrates the asymptotic validity of the confidence intervals proposed in (13). The result accommodates the case of misspecification often present in applied econometric research. It is most useful under the additional slight undersmoothing condition that makes any misspecification bias vanish sufficiently fast. In particular, when $\mathcal{G}$ is in a s-dimensional ball on $\mathcal{Z}$ of finite diameter (a Hölder class of smoothness order s) then the condition simplifies to $n^{1/2} k^{-(\frac{1}{2}+\frac{s}{d})} \log(k) \to 0$, see also Belloni et al. (2015), Comment 4.3 for additional details. Note that such undersmoothing does in general not admit IMSE optimal k choices. Alternatively, bias-correction methods could be employed (Cattaneo et al., 2020).

Theorem 5.1 extends readily to alternative combinations of Neyman-orthogonal scores other than $rATE$ or Δ. In particular, the results for the $rATE$ can directly be applied to any convex combination of conditional average treatment effects as long as the weights are either (i) deterministic sequences (relative to J) or (ii) can be estimated at the same rate as π_t. This can be useful when comparing heterogeneity of a given selection mechanism to alternative, hypothetical (estimated or true) allocation policies different from random

selection as considered in this paper even when there are potentially many different treatments available.

5.2 Convergence Rates when J is large: Examples

In this section we provide some basic examples and intuition about the properties of probability and nuisance function estimation when J is large and how this relates to the machine learning bias Assumptions B.3/C.3. We first discuss the necessity of Assumption A.3 and the consequences for the unconditional probability estimates. We then show how these properties translate into different convergence rates for propensity scores and potential outcome means under simplified parametric assumptions. We then discuss the explicit requirements for the machine learning bias Assumptions B.3/C.3 for the flexible high-dimensional nuisance parameter case using Lasso methodology under approximate sparsity and many treatments.

5.2.1 Large J and Assumption A.3

Consider the second part of Assumption A.3: If J is large, then $J\pi_t \lesssim 1$ is a necessary requirement. Because if the product diverges, $J\pi_t > 1$ causes a contradiction with the summability constraint $\sum_{t\neq 0} \pi_t = 1 - \pi_0$. In principle, one could allow for some $t \neq 0$ such that $J\pi_t = o(1)$. However, restricting A.3 to hold for all $t \neq 0$ is without loss of generality as otherwise it would be asymptotically equivalent to a regime where only a smaller subset of treatments $J' < J$ obey A.3. Thus, all further assumptions and rates would be equivalent with J being replaced by the new number of asymptotically relevant treatments J'.

5.2.2 Superefficiency of Unconditional Probability Estimators

(Local) superefficiency of frequency estimators $\hat{\pi}_t = \frac{1}{n}\sum_{i=1}^n D_{t,i}$ and other nuisance parameters is not a new discovery and has been exploited and discussed in different places in the literature, see e.g. Stoye (2009). For general $P(D_{t,i} = 1) = \pi_t$, note that $V[D_{t,i}] = \pi_t(1-\pi_t)$. Hence, $\sqrt{n}(\hat{\pi}_t - \pi_t) \xrightarrow{d} \mathcal{N}(0, \pi_t(1-\pi_t))$ which implies that $|\hat{\pi}_t - \pi_t| \lesssim_P (n/\pi_t)^{-1/2} \lesssim (nJ)^{-1/2}$

due to Assumption A.3. Thus under the many treatments regime $J \to \infty$, the frequency estimator is superefficient, i.e. converges at a quicker rate than $n^{-1/2}$.

5.2.3 Superefficiency of Parametric Propensity Scores

Superefficiency of $\hat{\pi}_t$ spills over to frequency-based/parametric estimators of propensity scores. For example, consider the case where $\mathcal{X}$ is discrete (finite-dimensional) with $f_x := P(X_i = x) > 0$ for all $x \in \mathcal{X}$. Consider a simple frequency-based estimator for the treatment propensity t as

$$\hat{e}_t(x) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i = x) + \mathbb{1}\left(\sum_{i=1}^{n} \mathbb{1}(X_i = x) = 0\right)\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i = x) D_{t,i}.$$

where the additional indicator in the denominator assures existence. By standard arguments, it follows that, for each $x \in \mathcal{X}$, $\sqrt{n}(\hat{e}_t(x) - e_t(x)) \xrightarrow{d} \mathcal{N}(0, e_t(x)(1 - e_t(x)/f_x))$ which implies that $|\hat{e}_t(x) - e_t(x)| \lesssim_P (n/e_t(x))^{-1/2} \lesssim (nJ)^{-1/2}$ due to Assumption A.3. Thus, the frequency-based finite-dimensional/parametric propensity score has the same superefficiency property as the unconditional frequency estimator.

5.2.4 Slower convergence of Parametric Mean Functions

Parametric estimators of potential outcome means, however, are not superefficient. On the contrary, convergence rates are generally slower under the many treatments regime. For example, consider a parametric frequency-based estimator similar to the one for the propensity score:

$$\hat{\mu}_t(x) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i = x) D_{t,i} + \mathbb{1}\left(\sum_{i=1}^{n} \mathbb{1}(X_i = x) D_{t,i} = 0\right)\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i = x) D_{t,i} Y_i$$

where $\mathcal{X}$ is again assumed to be discrete (finite-dimensional) with $f_x := P(X_i = x) > 0$ for all $x \in \mathcal{X}$. Without loss of generality, assume that $E[(Y_i(t) - \mu_t(X_i))^2 | X_i = x] = \sigma^2$ (Assumption B.1/C.1 would suffice as well). Again, by standard arguments, we have that, for all $x \in \mathcal{X}$, $\sqrt{n}(\hat{\mu}_t(x) - \mu_t(x)) \xrightarrow{d} \mathcal{N}(0, \sigma^2/e_t(x)f(x))$ which implies that $|\hat{\mu}_t(x) - \mu_t(x)| \lesssim_P (ne_t(x))^{-1/2} \lesssim (n/J)^{-1/2}$ due to Assumption A.3. Thus, the estimator converges at a slower than parametric rate. In fact, we have that, for any $t \neq 0$,

$s_{n,t,2} = Jm_{t,n,2} = \sqrt{J}s_{0,n,2}$. Thus, the corresponding components in the machine learning bias assumptions B.3/C.3 will be of equal rate in the parametric case.

5.2.5 Convergence for High-dimensional Sparse Nuisance Functions

Here we provide some intuition regarding the use of nuisance function estimation using (frequency-based) Lasso in high-dimensional approximately sparse models. In particular, we say potential outcome means are generated by $\mu_t(x) = x'\theta_0 + r_{\mu_t}(x)$ where $x \in \mathcal{X} \subseteq \mathbb{R}^{p_{\mu_t}}$ (and equivalently for $e_t(x)$ with logistic link). p_{μ_t} denotes the number of available regressors that is allowed to be high-dimensional and grow with n. Assume that the typical regularity conditions for Lasso hold as in Semenova and Chernozhukov (2021), Lemma B.1. Denote $s^*_{e_t}$ and $s^*_{\mu_t}$ as the corresponding sparsity indices that obey these assumptions. For simplicity, let the number of available regressors and sparsity indices coincide for propensity score and potential outcome estimation, i.e. $s^*_{e_t} = s^*_{\mu_t} \equiv s^*$ and $p_{e_t} = p_{\mu_t} \equiv p$. For the machine learning bias for the $nATE$ in Assumption B.3 we then conjecture that

$$
\begin{aligned}
B_n^{[nATE]} &\lesssim \sqrt{nk}\sqrt{\frac{s^* \log(p)}{n}}\left(J\sqrt{\frac{s^* \log(p)}{nJ}} + \sqrt{\frac{s \log(p)}{nJ}}\right) \\
&= s^*\sqrt{\frac{kJ \log(p)^2}{n}}
\end{aligned}
$$

based on the same argument as for the parametric frequency-based estimation above. Thus $B_n^{[nATE]} = o(1)$ requires that sparsity indices have to obey

$$
s^* = o\left(\sqrt{\frac{n}{kJ \log(p)^2}}\right).
$$

For $rATE/\Delta$, it is similarly required that

$$
s^* = o\left(\sqrt{\frac{n}{kJ^2 \log(p)^2}}\right).
$$

Thus, the sparsity conditions for $rATE/\Delta$ are stronger than for the $nATE$ in the many treatments regime. This again reflects the increased variability due to the different weighting schemes between the decomposition parameters as the $nATE$ weights minimize

variance as discussed in Section 5.1. Comparing the rate requirement for the $nATE$ to the one in Semenova and Chernozhukov (2021), we find that s^* here must be slower by a factor of $\sqrt{J}$ compared to their Lemma B.1. This is the price that has to be paid for the expanding set of nuisance parameters when estimating treatment propensities and potential outcome means for each treatment level separately instead of imposing the binary treatment structure to begin with. Moreover, the nonparametric heterogeneity analysis adds an additional factor of $\sqrt{k}$ to the sparsity requirements compared to standard double machine learning estimation of the unconditional binary ATE in Chernozhukov et al. (2018) that only requires $s = o(\sqrt{n/\log(p)^2})$ under equivalent assumptions. An analogous derivation can be conducted for Λ_n as well. Note that the given sparsity assumption here is for each treatment selection probability separately. In practice, we might want to impose some (group-based) sparsity across treatments to improve estimation when many treatment are available. In this case, convergence rates can be improved depending on the degree of total complexity of the propensity scores (Farrell, 2015). We leave an extension along these lines for future work.

6 Monte Carlo Study

In this section, we analyze the finite sample performance of the analytical confidence bounds proposed in Section 4. In particular, we evaluate the empirical coverage rates of the corresponding confidence intervals in a setup with heterogeneous effective treatment probabilities for all the decomposition parameters. We consider the case of three effective treatment levels and a univariate linear model for the heterogeneity analysis using different sample sizes and total number of confounding variables. In particular, in the final step, we regress the estimated pseudo outcomes on a single confounder and evaluate the coverage rates for the parameters of this linear predictor. We consider two ways to estimate the nuisance parameters (i) correctly specified parametric models and (ii) double machine learning estimators. For the latter we apply 2-fold cross-fitting using ℓ_1-regularized linear regression for the outcome models as well as ℓ_1-regularized multinomial logistic regression for the propensity scores. Tuning parameter selection is done via 5-fold cross-validation.

The true models satisfy the necessary sparsity assumptions required for high-quality approximation of the machine learning methods (Belloni & Chernozhukov, 2013; Farrell, 2015; Belloni, Chernozhukov, & Wei, 2016). For more details on the designs please consider Appendix B.7.

Table 2 contains the coverage rates of the confidence intervals based on (13) using correctly specified parametric models at a significance level of 5%. All results are very close to their nominal coverage rate. For the smallest $n = 1000$ and $k = 100$, there is undercoverage for β of 5 percentage points for the $rATE$ which largely vanishes for $n = 5000$. For the other parameters, there are no relevant size distortions.

Table 2: Monte Carlo Simulation: Results (Parametric Model)

(a) n = 1000

		$rATE$	$nATE$	Δ
k = 10	α	0.9460	0.9456	0.9524
	β	0.9444	0.9488	0.9470
k = 100	α	0.9468	0.9488	0.9518
	β	0.8974	0.9434	0.9474

(b) n = 5000

		$rATE$	$nATE$	Δ
k = 10	α	0.9486	0.9480	0.9482
	β	0.9448	0.9528	0.9570
k = 100	α	0.9506	0.9508	0.9508
	β	0.9380	0.9464	0.9458

The table entries contain the coverage rates under the null hypothesis for the parameters (α, β) of the linear predictor for different number of regressors (k), sample sizes (n) and decomposition parameters $rATE$, $nATE$ and Δ. The nominal coverage rate is 95%. All results are based on 5000 simulations.

Table 3 contains the coverage rates of the confidence intervals based on (13) using double machine learning at a significance level of 5%. For $rATE(x)$ and $nATE(x)$ all results are very close to the nominal coverage rate. For $\Delta(x)$, there is some undercoverage for the intercept α by 1.7 to 9.6 percentage points which increases in the number of parameters and decreases with the sample size. The slope parameter β is accurate for any sample or regressor set size. Overall the inference based on the asymptotic approximation in (13) seems to be reliable in finite samples.

7 Applications

7.1 Smoking and Birth Weight (Scenario 1)

The detrimental effect of smoking on birth weight and its economic costs are well documented (see e.g. Almond et al., 2005; Abrevaya, 2006; Almond & Currie, 2011, and

Table 3: Monte Carlo Simulation: Results (Double Machine Learning)

(a) n = 1000

		$rATE$	$nATE$	Δ
k = 10	α	0.9472	0.9494	0.8964
	β	0.9456	0.9452	0.9430
k = 100	α	0.9438	0.9420	0.8538
	β	0.9486	0.9476	0.9542

(b) n = 5000

		$rATE$	$nATE$	Δ
k = 10	α	0.9516	0.9500	0.9326
	β	0.9498	0.9506	0.9478
k = 100	α	0.9512	0.9508	0.9204
	β	0.9534	0.9520	0.9534

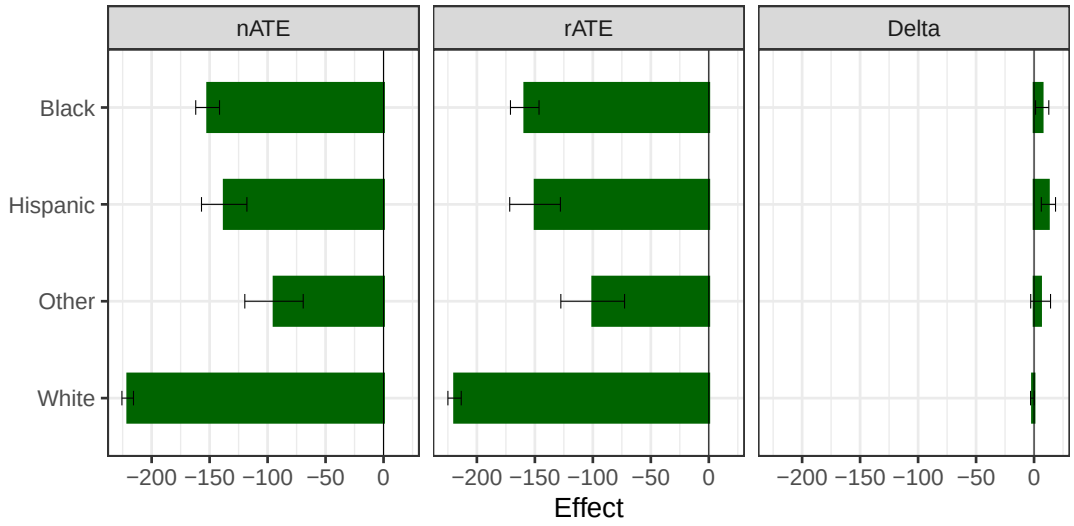
The table entries contain the coverage rates under the null hypothesis for the parameters (α, β) of the linear predictor for different number of regressors (k), sample sizes (n) and decomposition parameters $rATE$, $nATE$ and Δ. The nominal coverage rate is 95%. Results are based on 5000 simulations.

references therein). Beyond the standard average effects it is important to understand the heterogeneous effects to e.g. identify for which subgroups interventions to reduce smoking during pregnancy would be most beneficial. Abrevaya (2006) documents that the negative effect of smoking is less pronounced for black compared to white mothers in a standard subgroup analysis. A variety of papers analyze heterogeneous effects of smoking as a function of mother's age (Abrevaya, Hsu, & Lieli, 2015; Lee et al., 2017; Zimmert & Lechner, 2019; Fan et al., 2022). They all document increasingly negative effects with higher age. The aforementioned studies consider "smoking yes/no" as the binary treatment. Cattaneo (2010) notes that smoking is not a homogeneous treatment, but that the negative effects become more extreme for higher intensities of smoking. Thus, the binary indicator "smoking" represents only an aggregation of smoking intensities which directly affect birth weight. This corresponds to *Scenario 1.* We investigate whether the heterogeneous effects documented in the literature can be at least partly explained by different smoking intensities of different groups.

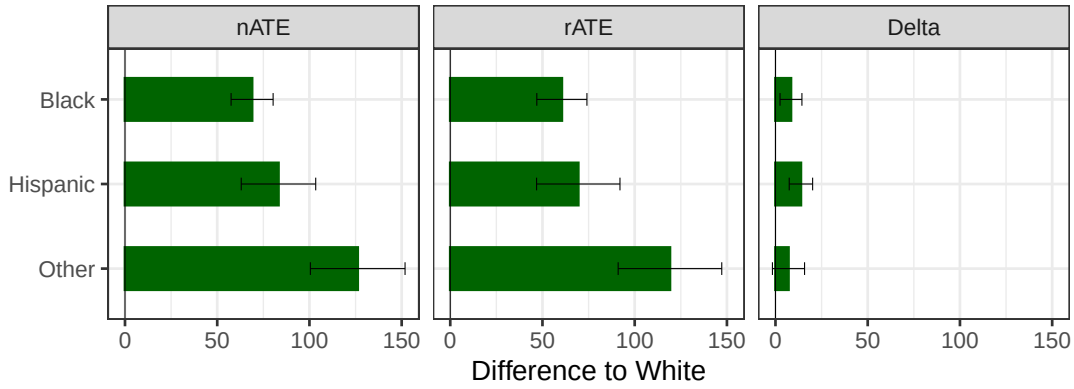
We analyze the dataset of Almond et al. (2005) used by Cattaneo (2010) with five intensities of smoked cigarettes per day as the effective treatment $T_i \in \mathcal{T} = \{0, 1-5, 6-10, 11-15, 16-20, >20\}$, the binary indicator defined as $D_i = \mathbb{1}(T_i > 0)$, the outcome Y_i being birth weight in gram, and the confounders X_i including age, education, ethnicity, and marital status of mother and father as well as health indicators and pregnancy history of the mother.[6] The dataset comprises 511,940 observations after removing the 0.1% of the observations with missing values in relevant variables and 52 confounders. The

[6]We thank Matias Cattaneo for sharing the full data. A random subsample is available on his GitHub repository.

Figure 3: Heterogeneous effects and decomposition by ethnicity



(a) Subgroup effects for ethnicity



(b) Effect heterogeneity with white as benchmark

Note: Point estimates of the decomposition parameters with 95%-confidence interval.

nuisance parameters are estimated with 2-fold cross-fitting using an ensemble learner of the unconditional mean, Random Forests, Lasso and Ridge regression with 2-fold cross-validated weights. For the propensity scores, we use logistic Lasso and Ridge.

Smoking behavior differs along the heterogeneity variables ethnicity and age showing that white and older smoking mothers smoke more heavily.[7] Combined with the result of Cattaneo (2010) that different smoking intensities have different effects, this suggests that at least part of the heterogeneity could be explained by different smoking intensities.

Figure 3 contains the result of the decomposition for the heterogeneity variable "ethnicity". The upper panel shows the decomposition for each subgroup. It is obtained by running an OLS regression of the estimand specific pseudo-outcome on a set of four

[7]Appendix B.8 and in particular Figure B.4 provides the smoking distributions by heterogeneity variables.

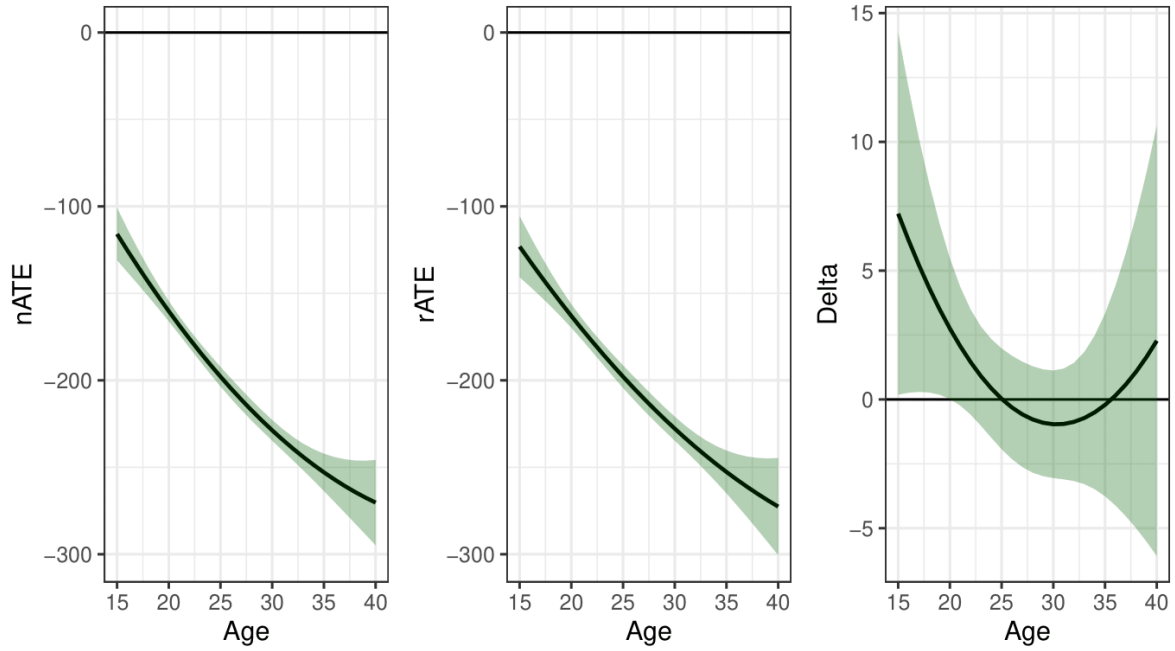
dummy variables indicating ethnicity of the mother without a constant. The standard errors are then adjusted as described in Section 4. The $nATE$ in the left part corresponds to standard subgroup analysis. Like previous studies, we find that smoking reduces the birth weight of newborns more for white women than for Blacks, Hispanics and others. Given that smoking is a binarized treatment, it is not clear how much is really effect heterogeneity and how much is driven by the fact that subgroups differ in their smoking intensity. The decomposition term $rATE$ fixes the intensity of smoking for all subgroups at the population level. It provides the subgroup specific effect of smoking if all groups had the same smoking intensity. Under this harmonized smoking intensity the negative average effect of smoking is smaller for white women and larger for the others. Δ in the right graph quantifies the difference between $nATE$ and $rATE$. It shows relatively small differences suggesting that different smoking intensities are not the main driver of the differences between white mothers and the other groups. However, they are also not negligible as the lower panel of Figure 3 shows. It quantifies the heterogeneous effects by subtracting the effects for white mothers from the other three groups. We observe that a significant portion of the difference between black/hispanic mothers and white mothers is driven by different smoking intensities. For black vs. white mothers the difference in the $nATE$ is 69 gram of which 12% are due to different smoking intensities ($\Delta = 8$). For hispanic vs. white mothers it explains around 17% ($\Delta = 14$).

Figure 4 depicts the heterogeneity analysis along age. We use B-splines as basis functions of age. We select the nodes and order via leave-one-out cross-validation for each parameter and apply the most flexible/low-bias model for all parameters to ensure that the $rATE$ and Δ curves add up to the $nATE$ curve. The left panel of Figure 4 replicates the well-established findings of previous papers that the $nATE$ is much smaller for younger mothers than for older mothers.

In the extreme case where different smoking intensities would fully explain the heterogeneous $nATE$, we would see a flat $rATE$ curve in the middle graph. However, we only observe that the effect of teenage mothers would be more negative if we harmonize smoking intensity over all age groups.

Overall, only a relatively small part of the heterogeneous effects of the binarized

Figure 4: Effect Heterogeneity by Age



Notes: B-spline estimated decomposition parameters with 95%-confidence interval.

smoking indicator can be attributed to different smoking intensities and the larger part seems to be driven by different age groups actually being affected differently.

7.2 Job Corps (Scenario 2)

We illustrate Scenario 2 of Figure 2 with an evaluation of the Job Corps (JC) program. JC operates since 1964 and is the largest training program for disadvantaged youth aged 16-24 in the US (see Schochet et al., 2001, 2008, for a detailed description). The roughly 50,000 participants per year receive an intensive treatment as a combination of different components like academic education, vocational training, and job placement assistance. Participants plan their educational and vocational curricula together with counselors. This means that although the variable "access to JC" is a binary indicator, different versions of JC participation are conceivable. Heterogeneous effects might thus be driven by different effectiveness of JC for different groups, by different tailoring of the curriculum, or a combination thereof.

We investigate this based on data from an experiment in 1994-1996 (Schochet, Burghardt,

& McConnell, 2019).[8] This experiment is basis of a variety of studies looking at different aspects of JC. Many of them report gender differences in the effectiveness of the programs with women benefiting less than men from access to JC (e.g. Schochet et al., 2001, 2008; Flores et al., 2012; Eren & Ozbelik, 2014; Strittmatter, 2019). One potential explanation for this finding is that men and women focus on average on different vocational training within JC. In particular men receive more often training for higher paying craft jobs, while women focus more often on training for the service sector (Quadagno & Fobes, 1995; Inanc, Needels, & Berk, 2017).[9] We apply our decomposition method to investigate this potential explanation of the gender gap in program effectiveness.

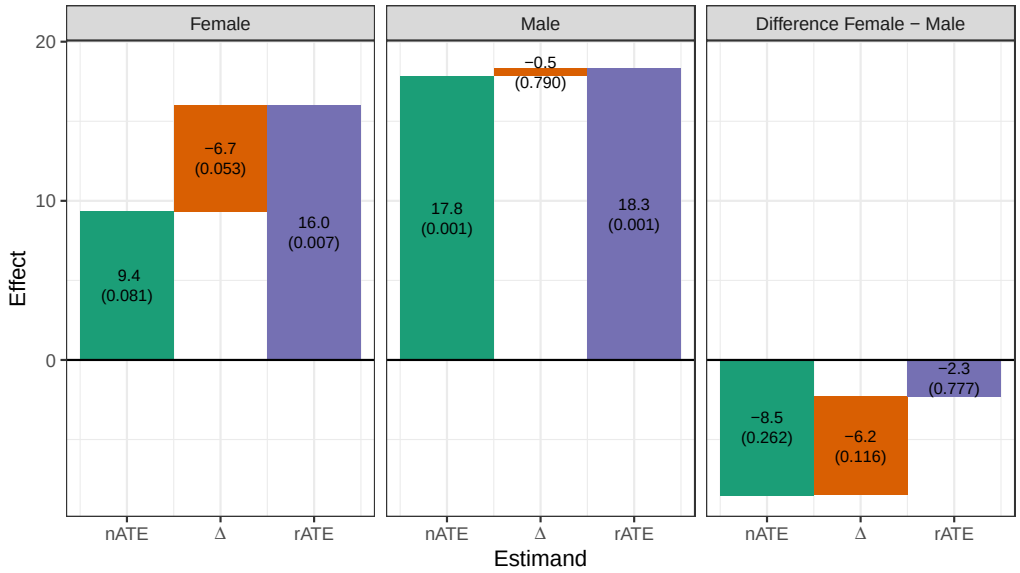
We analyze the intention to treat effect (ITT) of the binary variable indicating random access to JC (D_i) on weekly earnings four years after random assignment (Y_i). We consider 11 versions of the effective treatment (T_i): (i) *No JC* if eligible individuals did not participate (non-compliers), (ii) *JC without vocational training* if eligible individuals entered JC but did not receive vocational training, (iii-ix) training for jobs in the clerical, health, auto mechanics, welding, electrical/electronics, construction, or food sector, (x) other vocational training, (xi) training for multiple sectors.

Nuisance parameters are estimated with the same ensemble as in Section 7.1 using 5-fold cross-fitting. We control for 55 covariates that include pre-treatment information about labor market history, socio-economic characteristics, education, health, crime, and JC related variables. These control variables overlap mostly with those of Flores et al. (2012) who also employ an unconfoundedness strategy. Considering second-order interactions results in a total of 1428 variables after screening for nearly empty cells (less than 1% observations) and nearly perfectly correlated variables (correlation higher than 0.99). In total we work with a sample of 9,708 observations.

The unconditional $nATE$, corresponding to the ITT of eligibility for JC on monthly earnings, is estimated at \$14.2 (S.E. 3.8), which is an increase of 7% in line with previous studies. The unconditional $rATE$ is larger (\$17.4, S.E. 4.1) suggesting that hypothetical random allocation of the curricula would yield higher average outcomes compared to the actual assignment. However, the unconditional difference Δ is insignificant ($\$ - 3.1$,

[8]The data is available as public use file via https://doi.org/10.3886/E113269V1.
[9]Appendix B.9 and in particular Figure B.7 provides the distribution of trainings by gender.

Figure 5: Effect heterogeneity and decomposition by gender



Notes: The numbers in the bar show the point estimate and the p-value in parentheses.

S.E. 1.8). This suggests that, on average, the selection of versions is not statistically distinguishable from random allocation.

Figure 5 depicts the decomposition of the gender specific effects. We observe that the effect for women with the actual composition of vocational training ($nATE$) is not significant at $\alpha = 0.05$, but under the hypothetical treatment composition of the population would show a clear positive effect ($rATE = \$16.0$). The gender gap in effectiveness basically disappears when both groups receive the same hypothetical mix of vocational training. The right part of Figure 5 suggests that 73% of the gender gap in the effectiveness of JC is due to different training curricula. This means that the worse than average performance of the assignment mechanism seen in the unconditional parameters is mostly driven by women. While the assignment to vocational training for men is as well targeted as random assignment, for women it is even worse. This indicates that there is room for improvement to target vocational training in general and for women in particular. Our results suggest that removing the worse than random targeting of vocational training for women could decrease the gender gap in the effectiveness of access to JC.

8 Concluding Remarks

The method proposed in this paper provides a practical way of decomposing effect heterogeneity obtained from analyzing a binary treatment indicator that does not coincide with the effective multi-valued treatment. The approach likely extends to other causal parameters and identification strategies such as continuous effective treatments, selection on unobservables/instrumental variables, or mediation analysis. It would also be interesting to see whether the ideas could be further developed to find the most relevant dimensions of effective treatments for cases with multiple treatment versions instead of requiring the researcher to manually specify them.

The conceptual and empirical results highlight that potential treatment heterogeneity underlying the analyzed binary indicator should be taken more seriously and explicitly discussed in applications, especially when interpreting heterogeneous effects. The decomposition provides one principled way to do this. It requires to observe the effective treatment. Thus, data collection can anticipate the goal of better understanding treatment heterogeneity by recording effective treatment information beyond a binary indicator. Furthermore, the decomposition shows that reducing the analysis to such binary indicators, while facilitating the analysis, comes at the cost of a more intricate interpretation of empirical results.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*, 465–503.

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics*, *21*(4), 489–519.

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

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The costs of lower birth weight. *The Quarterly Journal of Economics*, *120*(3), 1031–1083.

Almond, D., & Currie, J. (2011). Human capital development before age ve. In Handbook of labor economics (Vol. 4, pp. 1315 1486). Elsevier.

Andresen, M. E., & Huber, M. (2021). Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction. The Econometrics Journal, 24(3), 536 558.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal e ects in models with variable treatment intensity. Journal of the American Statistical Association, 90(430), 431 442.

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal e ects. Proceedings of the National Academy of Sciences, 113(27), 7353 7360.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3 32.

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

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2), 521 547.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. Journal of Econometrics, 186(2), 345 366.

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. Journal of Business & Economic Statistics, 34(4), 606 619.

Buhl-Wiggers, J., Kerwin, J., Muñoz, J. S., Smith, J., & Thornton, R. (2022). Some children left behind: variation in the e ects of an educational intervention. Journal of Econometrics doi: 10.1016/j.jeconom.2021.12.010

Cattaneo, M. D. (2010). E cient semiparametric estimation of multi-valued treatment e ects under ignorability. Journal of Econometrics, 155(2), 138 154.

Cattaneo, M. D., Farrell, M. H., & Feng, Y. (2020). Large sample properties of partitioning-based series estimators. Annals of Statistics, 48(3), 1718 1741.

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016). Interpreting

regression discontinuity designs with multiple cutoffs. *Journal of Politics*, *78*(4), 1229–1248.

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.

Cole, S. R., & Frangakis, C. E. (2009). The consistency statement in causal inference. *Epidemiology*, *20*(1), 3–5.

Curth, A., & van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th international conference on artificial intelligence and statistics* (Vol. 130, pp. 1810–1818). PMLR.

Davis, J. M. V., & Heller, S. B. (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *The Review of Economics and Statistics*, *102*(4), 664–677.

Eren, O., & Ozbelik, S. (2014). Who benefits from Job Corps? A distributional ananlysis of an active labor market program. *Journal of Applied Econometrics*, *29*(4), 586–611.

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.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, *189*(1), 1–23.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Review of Economics and Statistics*, *94*(1), 153–171.

Heckman, J. J. (2020). Epilogue: Randomization and social policy evaluation revisited. In F. Bédécarrats, I. Guérin, & F. Roubaud (Eds.), *Randomized control trials in the field of development: A critical perspective* (pp. 304–330). Oxford University Press.

Heiler, P. (2022). Estimating Heterogeneous Bounds for Treatment Effects under Sample Selection and Non-response. *arXiv:2209.04329*. Retrieved from http://arxiv.org/abs/2209.04329

Heiler, P., & Kazak, E. (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics*, *222*(2), 1083–1108.

Heiler, P., & Taylor, L. (2022). Nonparametric estimation for categorical responses with small frequencies. *Working Paper*.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, *22*(3), 368–377.

Hong, H., Leung, M. P., & Li, J. (2020). Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, *23*(1), 32–47.

Hotz, V. J., Imbens, G. W., & Klerman, J. A. (2006). Evaluating the differential dffects of alternative welfare-to-work training aomponents: A reanalysis of the California GAIN program. *Journal of Labor Economics*, *24*(3), 521–566.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*(1-2), 241–270.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, *31*(5), 1600–1635.

Imai, K., & Li, M. L. (2021). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1923511

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

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

Inanc, H., Needels, K., & Berk, J. (2017). *Gender segregation in training programs and the wage gap* (Tech. Rep. Nos. Cambridge, NJ: Mathematica Policy Research).

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497*. Retrieved from http://arxiv.org/abs/2004.14497

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*, 1229–1245.

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, *78*(6), 2021–2042.

Knaus, M. C. (2022). Double machine learning based program 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. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner & E. Pfeiffer (Eds.), *Econometric evaluation of labour market policies* (pp. 43–58). Heidelberg: Physica.

Lechner, M. (2002). Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *The Review of Economics and Statictics*, *84*, 205–220.

Lee, S., Okui, R., & Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, *32*(7), 1207–1225.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, *113*(521), 390–400.

Ma, X., Sasaki, Y., & Wang, Y. (2022). Testing limited overlap.

Ma, X., & Wang, J. (2020). Robust Inference Using Inverse Probability Weighting. *Journal of the American Statistical Association*, *115*(532), 1851–1860.

Marshall, J. (2016). Coarsening bias: How coarse treatment measurement upwardly biases instrumental Variable Estimates. *Political Analysis*, *24*(2), 157–171.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, *79*(1), 147–168.

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

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Petersen, M. L. (2011). Compound treatments, transportability, and the structural causal model: The power and simplicity of causal graphs. *Epidemiology*, *22*(3), 378–381.

Quadagno, J., & Fobes, C. (1995). The welfare state and the cultural reproduction of gender: Making good girls and boys in the Job Corps. *Social Problems*, *42*(2), 171–190.

Richardson, T., & Robins, J. M. (2013). Single World Intervention Graphs (SWIGs): Unifying the Counterfactual and Graphical Approaches to Causality – Presentation. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Rothe, C. (2017). Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap. *Econometrica*, *85*(2), 645–660.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, *164*(1), 60–72.

Schochet, P. Z., Burghardt, J., & Glazerman, S. (2001). *National job corps study: The impacts of job jorps on participants' employment and related outcomes* (Tech. Rep.). Princeton, NJ: Mathematica Policy Research Inc.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does job corps work? Impact

findings from the national job corps study. *American Economic Review*, *98*(5), 1864–1886.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2019). Replication data for: Does job corps work? Impact findings from the national job corps study. *Inter-university Consortium for Political and Social Research (ICPSR)*.

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

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, *77*(4), 1299–1315.

Strittmatter, A. (2019). Heterogeneous earnings effects of the job corps by gender: A translated quantile approach. *Labour Economics*, *10760*.

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology*, *20*(6), 880–883.

VanderWeele, T. J., & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, *1*(1), 1–20.

Vazquez-Bare, G. (2022). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*. doi: 10.1016/j.jeconom.2021.10.014

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.

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779*. Retrieved from http://arxiv.org/abs/1908.08779

Appendices

A Proof of Theorem 5.1

A.1 Preliminaries

The proof is structured as follows: First we provide some auxiliary results. Then we derive the asymptotically linear representation of the best linear predictor and show its asymptotic normality using the true variance covariance matrix. The necessary derivations to replace the true variance with the estimated sample counterpart are provided in Supplementary Appendix B.6. For reference, we refer with BCCK to Belloni et al. (2015) and with SC to Semenova and Chernozhukov (2021). For the following, we use empirical process notation

$$E_n[X_i] := \frac{1}{n}\sum_{i=1}^{n} X_i, \quad G_n[X_i] := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E[X_i]).$$

Recall that $D_i = \sum_{t\neq 0} D_{t,i}$ and $D_i Y_i = \sum_{t\neq 0} D_{t,i} Y_i$. Conditional independence implies that

$$E[Y_i D_i | X_i] = \sum_{t\neq 0} e_t(X_i)\mu_t(X_i), \quad E[\mu_t(X_i) D_i | X_i] = \sum_{t\neq 0} e_t(X_i)\mu_t(X_i).$$

Recall that $D_{t,i} D_{s,i} = 0$ for $s \neq t$.

A.2 Machine Learning Bias

In the following, we verify the small bias Assumption 3.5 in SC for our moment functions evaluated at the true π. Let $u_n = o(1)$ such that with probability of at least $1 - u_n$, for all $f \in [K]$, the cross-fitted $\hat{\eta}_f$ belongs to a shrinking neighborhood $\mathcal{H}_n$ around η. We show that, uniformly over $\mathcal{H}_n$, the moment functions for $nATE$, $rATE$, and Δ satisfy

$$B_n = \sqrt{n} \sup_{\hat{\eta}\in\mathcal{H}_n} ||E[b(Z_i)(\psi(\hat{\eta}, \pi) - \psi(\eta, \pi))]|| = o(1)$$

$$\Lambda_n = \sup_{\hat{\eta}\in\mathcal{H}_n} E[||b(Z_i)(\psi(\hat{\eta}, \pi) - \psi(\eta, \pi))||^2]^{1/2} = o(1).$$

For the proof the followings expectations are all used omitting prefix $\sup_{\hat{\eta}\in\mathcal{H}_n}$ when it does not cause confusion. We will make use of a general decomposition of the AIPW-type moment functions: For general binary D_i, any Y_i, and $\eta = (\mu_i, e_i)$ we define

$$\psi_i(\eta) = \frac{D_i(Y_i - \mu_i)}{e_i} + \mu_i.$$

Decomposing the function evaluated at two points yields

$$\begin{aligned}\psi_i(\hat{\eta}) - \psi_i(\eta) &= (\hat{\mu}_i - \mu_i)\left(1 - \frac{D_i}{e_i}\right) - (\hat{e}_i - e_i)(Y_i - \mu_i)\frac{D_i}{e_i\hat{e}_i} + (\hat{\mu}_i - \mu_i)(\hat{e}_i - e_i)\frac{D_i}{\hat{e}_i e_i} \\ &\equiv (a.1) - (a.2) + (a.3).\end{aligned}$$

For the following we will ignore the "control" part $\psi_i^{[0]}(W_i, \eta)$ in the moment functions as this is covered by the standard potential outcome case in SC. The rates in the following are not affected. For the following it is essential to note that, due to cross-fitting, $\hat{e}_t(X_i)$ and $\hat{\mu}_t(X_i)$ only depend on i through X_i. Thus, evaluating expectations depending on $\hat{\eta}$, we omit the explicit conditioning set on the cross-fitted fold for convenience as in SC. The unconditional convergence then follows from Chernozhukov et al. (2018), Lemma 6.1. Note also that, conditional on X_i, any measurable function of Z_i is known. We first provide some auxiliary results in what follows.

A.2.1 Auxiliary results

(H.1): $||\hat{\gamma}_t - \gamma_t||$ bound

$$||\hat{\gamma}_t - \gamma_t|| \leq ||E_n[b_i(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))]|| - ||E_n[b_i\psi_i^{[t]}(\eta) - E[b_i\psi_i^{[t]}(\eta)]]||$$

Using Markov's inequality and Cauchy-Schwarz together with the definition of the moment function yields for the first term

$$\begin{aligned}
||E_n[b_i(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))]|| &\lesssim_P E[||b_i(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))||] \\
&\lesssim_P E[||b_i||^2]^{1/2} E[(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))^2]^{1/2} \\
&\leq \sqrt{k} E[(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))^2]^{1/2} \\
&\lesssim_P \sqrt{kJ}(m_{t,n,2} + J s_{t,n,2})
\end{aligned}$$

as

$$\begin{aligned}
&E[(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))^2] \\
&\quad \leq 4E\Bigg[(\hat{\mu}_t(X_i) - \mu_t(X_i))^2(1 - D_{t,i}/e_t(X_i))^2 + (\hat{e}_t(X_i) - e_t(X_i))^2 \varepsilon_i(t)^2 \frac{D_{t,i}}{e_t(X_i)^2 \hat{e}_t(X_i)^2} \\
&\qquad + (\hat{\mu}_t(X_i) - \mu_t(X_i))^2 (\hat{e}_t(X_i) - e_t(X_i))^2 \frac{D_{t,i}}{e_t(X_i)^2 \hat{e}_t(X_i)^2}\Bigg] \\
&\quad \lesssim_P \sup_{x \in \mathcal{X}} \frac{\pi_t}{e_t(x)} \pi_t^{-1} \Bigg(E[(\hat{\mu}_t(X_i) - \mu_t(X_i))^2] + \chi_{t,n}^2 \pi_t^{-2} E[(\hat{e}_t(X_i) - e_t(X_i))^2] \Bigg) \\
&\quad \lesssim_P J(m_{t,n,2}^2 + J^2 s_{t,n,2}^2)
\end{aligned}$$

For the second term we have

$$\begin{aligned}
||E_n[b_i \psi_i^{[t]}(\eta) - E[b_i \psi_i^{[t]}(\eta)]]|| &\lesssim_P E[||E_n[b_i \psi_i^{[t]}(\eta)] - E[b_i \psi_i^{[t]}(\eta)]||] \\
&\leq E[||E_n[b_i \psi_i^{[t]}(\eta)]||^2]^{1/2} \\
&= (E[\psi_i^{[t]}(\eta)^2 b_i' b_i]/n)^{1/2} \\
&= (E[E[\psi_i^{[t]}(\eta)^2 | X_i] b_i' b_i / n])^{1/2} \\
&= \left(E\left[\left(\frac{\sigma_t^2(X_i)}{e_t(X_i)} + \mu_t(X_i)^2 \right) b_i' b_i \right] / n \right)^{1/2} \\
&\lesssim \left(\left(\sup_{x \in \mathcal{X}} \frac{\pi_t}{e_t(x)} \pi_t^{-1} + 1 \right) E[b_i' b_i]/n \right)^{1/2} \\
&\lesssim \sqrt{\frac{Jk}{n}}
\end{aligned}$$

by conditional independence and bounded second moments for the potential outcomes. Overall we obtain

$$||\hat{\gamma}_t - \gamma_t|| \lesssim_P \sqrt{kJ}\left(n^{-1/2} + m_{t,n,2} + Js_{t,n,2}\right)$$

(H.2): $||\gamma_t||$ **rate**

$$||\gamma_t|| = ||E[b_i\psi_i^{[t]}(\eta)]|| = ||E[b_i\mu_t(X_i)]|| \lesssim \sup_{x\in\mathcal{X}}|\mu_t(x)|E[||b_i||] \lesssim \sqrt{k}$$

(H.3): $||E_n[b_i(\psi_i(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))]||$ **rate**

$$\begin{aligned}
||E_n[b_i(\psi_i(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))]|| &= ||\sum_{t\neq 0} E_n[b_i\psi_i^{[t]}(\hat{\eta})]\left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0}\right)|| \\
&\lesssim_P J\sup_{t\neq 0}(||\gamma_t|| + ||\hat{\gamma}_t - \gamma_t||)|\hat{\pi}_t - \pi_t|(1 + n^{-1/2}) \\
&\lesssim_P J\sqrt{k}\sqrt{\frac{1}{nJ}} \\
&= \sqrt{\frac{Jk}{n}}
\end{aligned}$$

(H.4) Restatement of Lemma 6.2 from BCCK of the Rudelson (1999) LLN for Matrices Let $Q_1, \ldots, Q_n$ be a sequence of independent symmetric non-negative $k \times k$-matrix valued random variables with $k \geq 2$ such that $Q = E_n[E[Q_i]]$ and $||Q_i|| \leq M$ a.s., then for $\hat{Q} = E_n[Q_i]$

$$E[||\hat{Q} - Q||] \lesssim \frac{M\log k}{n} + \sqrt{\frac{M||Q||\log k}{n}}.$$

(H.5) γ**,** a_i**, and** a_ia_i' **rates** Define $A_n = E_n[a_ia_i']$ and $\hat{A}_n = E_n[\hat{a}_i\hat{a}_i']$. Recall the definitions:

$$\gamma = (\gamma_1, \ldots, \gamma_J),\ \gamma_t = E[b_i\tau_t(X_i)], \quad a_i = (a_i^{[1]}, \ldots, a_i^{[J]})$$

Thus

$$||\gamma|| \lesssim \sqrt{J} \sup_{t \neq 0} ||\gamma_t|| \lesssim \sqrt{Jk}$$

$$||\hat{\gamma} - \gamma|| \lesssim_P \sqrt{J} \sup_{t \neq 0} ||\hat{\gamma}_t - \gamma_t|| \lesssim_P \sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2})$$

$$E[||a_i||] \leq \sqrt{J} \sup_{t \neq 0} E[(a_i^{[t]})^2]^{1/2} \lesssim_P \sqrt{J} \sup_{t \neq 0} \sqrt{\pi_t} \lesssim 1$$

$$E[||\hat{a}_i - a_i||] \leq \sqrt{J} \sup_{t \neq 0} E[(\hat{a}_i^{[t]} - a_i^{[t]})^2]^{1/2} \lesssim \sqrt{J} \sup_{t \neq 0} \sqrt{\frac{\pi_t}{n}} \lesssim n^{-1/2}$$

as $\hat{a}_i^{[t]} - a_i^{[t]} \lesssim_P \pi_t|\hat{\pi}_0 - \pi_0| + \pi_0|\hat{\pi}_t - \pi_t|$. Equivalently, we have

$$||A_n|| \leq E_n[||a_i a_i'||] \lesssim_P E[||a_i||^2] \leq J \sup_{t \neq 0} \pi_t \lesssim 1$$

$$||\hat{A}_n - A_n|| \lesssim E_n[||(\hat{a}_i - a_i)a_i'||] \lesssim_P \sqrt{J} \sup_{t \neq 0} \left(\pi_t(\hat{\pi}_0 - \pi_0)^2 + (\hat{\pi}_t - \pi_t)^2 \right)^{1/2} \lesssim_P n^{-1/2}$$

We further have that

$$\begin{aligned}
\max_{1 \leq i \leq n} ||a_i a_i'|| &\leq \sqrt{J} \max_{1 \leq i \leq n} ||a_i a_i'||_1 \\
&\lesssim_P \sqrt{J} \max_{1 \leq i \leq n} \sum_{t \neq 0} (D_{t,i}(1 - \pi_0) - \pi_t + D_{0,i}\pi_t)^2 \\
&\lesssim_P \sqrt{J}(1 + J^{-2} + J^{-1}) \\
&\lesssim \sqrt{J}
\end{aligned}$$

and for the expectation

$$\begin{aligned}
||E[a_i a_i']|| &\leq \sqrt{J}||E[a_i a_i']||_1 \\
&\leq \sqrt{J} \sup_{t \neq 0} (E[(a_i^{[t]})^2] + \sum_{s \neq t,0} E[a_i^{[t]} a_i^{[s]}]) \\
&\leq \sqrt{J} \sup_{t \neq 0} (\pi_t + \sum_{s \neq t,0} \pi_t \pi_s) \\
&\lesssim J^{-1/2}
\end{aligned}$$

Now note that $a_i a_i'$ are symmetric, non-negative iid matrix valued random variables. Thus

using Rudelson's LLN (H.4) we obtain

$$E[||E_n[a_i a_i'] - E[a_i a_i']||] \lesssim \frac{J^{1/2}\log J}{n} + \sqrt{\frac{J^{1/2}||E[a_i a_i']|| \log J}{n}} \lesssim \sqrt{\frac{\log J}{n}}$$

as $J = o(n)$.

(H.6) Linearization of the unconditional weights:

$$\begin{aligned}\frac{\hat{\pi}_t}{\sum_{t\neq 0}\hat{\pi}_t} - \frac{\pi_t}{\sum_{t\neq 0}\pi_t} &= \frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0} \\ &= \frac{\hat{\pi}_t(1-\pi_0) - (1-\pi_0)\pi_t + (1-\pi_0)\pi_t - \pi_t(1-\hat{\pi}_0)}{(1-\pi_0)(1-\hat{\pi}_0)} \\ &= \frac{1-\pi_0}{1-\hat{\pi}_0}\frac{1}{(1-\pi_0)^2}E_n[([D_{t,i}-\pi_t](1-\pi_0) + [D_{0,i}-\pi_0]\pi_t)]\end{aligned}$$

and thus

$$\sqrt{n}\left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0}\right) = \frac{1-\pi_0}{1-\hat{\pi}_0}G_n[a_i^{[t]}]$$

where $a_i^{[t]} = (1-\pi_0)^{-2}(D_{t,i}(1-\pi_0) + D_{0,i}\pi_t - \pi_t)$.

(H.7) $rATE$ **Conditional mean error variance:** Recall that $\varepsilon_i = \psi_i(\eta,\pi) - E[\psi_i(\eta,\pi)|Z_i]$. The conditional mean error for the $rATE$ has finite second conditional moment:

$$\begin{aligned}E[\varepsilon_i^2|Z_i] &= (1-\pi_0)^{-2}\sum_{t\neq 0}\sum_{t'\neq 0}\pi_t\pi_{t'}' E[(\psi_i^{[t]}(\eta) - E[\psi_i^{[t]}(\eta)|Z_i])(\psi_i^{[t']}(\eta) - E[\psi_i^{[t']}(\eta)|Z_i])|Z_i] \\ &\lesssim_P \sum_{t\neq 0}\pi_t^2 E\left[\frac{\sigma_t^2(X_i)}{e_t(X_i)} + (\mu_t(X_i) - E[\mu_t(X_i)|Z_i])^2\middle| Z_i\right] \\ &\quad + \sum_{t\neq 0}\sum_{t'\neq 0}\pi_t\pi_{t'}E[(\mu_t(X_i) - E[\mu_t(X_i)|Z_i])(\mu_{t'}(X_i) - E[\mu_{t'}(X_i)|Z_i])|Z_i] \\ &\lesssim J\sup_{t\neq 0}\pi_t \sup_{x\in\mathcal{X}}\frac{\pi_t}{e_t(x)} + J^2\sup_{t\neq 0}\pi_t^2 \\ &\lesssim 1\end{aligned}$$

As $\sup_{t,x\in\mathcal{X}}\sigma_t^2(x) + \mu_t(x)$ is uniformly bounded by Assumptions A.2 and B.1/C.1.

(H.8) Effect of estimating π **for decomposition term** $\Delta = nATE - rATE$**:** Note

that due to the multiplicative structure of the decomposition we have that, for any η,

$$\begin{aligned}
&\psi_i^{[\Delta]}(\eta, \hat{\pi}) - \psi_i^{[\Delta]}(\eta, \pi) \\
&\quad = [\psi_i^{[nATE]}(\eta, \hat{\pi}) - \psi_i^{[rATE]}(\eta, \hat{\pi})] - [\psi_i^{[nATE]}(\eta, \pi) - \psi_i^{[rATE]}(\eta, \pi)] \\
&\quad = \sum_{t \neq 0} \psi_i^{[t,0]}(\eta) \left[\frac{\pi_t}{\sum_{t \neq 0} \pi_t} - \frac{\hat{\pi}_t}{\sum_{t \neq 0} \hat{\pi}_t} \right] \\
&\quad = -(\psi_i^{[rATE]}(\eta, \hat{\pi}) - \psi_i^{[rATE]}(\eta, \pi))
\end{aligned}$$

Thus we can obtain an analogous asymptotically linear representation for the estimator of Δ in what follows using the $rATE$ results with a simple sign flip which does not affect the rates of the approximation bounds in the following. Moreover, the leading term used for the normality approximation and its asymptotic variance will also be identical up to the sign flip.

(H.9) Error and error tail bounds For the $rATE$, the regression error by definition is a convex combination of centered moment functions

$$\varepsilon_i = \frac{\sum_{t \neq 0} \pi_t (\psi_i^{[t]}(\eta) - E[\psi_i^{[t]}(\eta) | Z_i])}{\sum_{t \neq 0} \pi_t}$$

and thus

$$|\varepsilon_i| \leq \sup_{t \neq 0} |\psi_i^{[t]}(\eta) - E[\psi_i^{[t]}(\eta) | Z_i]| \lesssim \sup_{t \neq 0} \left| \frac{\varepsilon_i(t) D_{t,i}}{e_t(X_i)} \right| + 1$$

almost surely due to the bounded conditional means. For the $nATE$, however, the propensity score weights yield

$$\varepsilon_i = \frac{\sum_{t \neq 0} D_{t,i} \varepsilon_i(t) + e_t(X_i) \mu_t(X_i) - E[D_{t,i} \varepsilon_i(t) + e_t(X_i) \mu_t(X_i) | Z_i]}{\sum_{t \neq 0} e_t(X_i)}$$

and thus

$$|\varepsilon_i| \lesssim \sup_{t \neq 0} |\varepsilon_i(t)| + 1$$

almost surely.

A.2.2 Machine Learning Bias: $nATE$

For the $nATE$ the $\psi_i(\eta)$-function has components

$$\mu_i = \frac{\sum_{t\neq 0} e_t(X_i)\mu_t(X_i)}{\sum_{t\neq 0} e_t(X_i)},\ e_i = \sum_{t\neq 0} e_t(X_i),\ D_i = \sum_{t\neq 0} D_{t,i},\ Y_i = \sum_t D_{t,i}Y_i(t)$$

and equivalently for $\psi(\hat{\eta})$. We now verify the rates in SC, Assumption 3.5: For B_n, we have that

$$\begin{aligned}
E[b_i(a.1)] &= E[b_i[\hat{\mu}_i - \mu_i](1 - E[D_i|X_i]/e_i)] \\
&= 0 \\
E[b_i(a.2)] &= E[b_i[\hat{e}_i - e_i](E[Y_iD_i|X_i] - \mu_i E[D_i|X_i])/(e_i\hat{e}_i)] \\
&= 0 \\
||E[b_i(a.3)]|| &\lesssim \sup_{x\in\mathcal{X}}(1 - e_0(x))^{-1} E[||b_i||^2]^{1/2} E[(\hat{e}_i - e_i)^2(\hat{\mu}_i - \mu_i)^2]^{1/2} \\
&\lesssim \sqrt{k} E[(\hat{e}_i - e_i)^{2h_2}]^{\frac{1}{2h_2}} E[(\hat{\mu}_i - \mu_i)^{2h_1}]^{\frac{1}{2h_1}} \\
&\lesssim \sqrt{k} J s_{0,n,2h_2}\left(\sup_{t\neq 0} s_{t,n,2h_1} + \sup_{t\neq 0, x\in\mathcal{X}} e_t(x) m_{t,n,2h_1}\right) \\
&\lesssim \sqrt{k} s_{0,n,2h_2}\left(J s_{n,2h_1} + m_{n,2h_1}\right)
\end{aligned}$$

for some $1/h_1 + 1/h_2 = 1$ by Hölder's inequality. The aggregate $\hat{\mu}_i$ rate follows from Assumption A.3 together with expanding

$$\begin{aligned}
\hat{e}_t(x)\hat{\mu}_t(x) - e_t(x)\mu_t(x) = e_t(x)(\hat{\mu}_t(x) - \mu_t(x)) + \mu_t(x)(\hat{e}_t(x) - e_t(x)) \\
+ (\hat{e}_t(x) - e_t(x))(\hat{\mu}_t(x) - \mu_t(x))
\end{aligned}$$

for all t. As potential outcome means are uniformly bounded by A.2, this yields that

$$E[(\hat{\mu}_i - \mu_i)^c]^{1/c} \lesssim J\left(\sup_{t\neq 0} s_{t,n,c} + \sup_{t\neq 0, x\in\mathcal{X}} e_t(x) m_{t,n,c}\right)$$

for any $c \geq 2$. Overall we have that

$$\begin{aligned} B_n^{[nATE]} &= \sqrt{n} \sup_{\hat{\eta} \in \mathcal{H}_n} ||E[b_i(\psi_i^{nATE}(\hat{\eta}) - \psi_i^{nATE}(\eta))]|| \\ &\lesssim \sqrt{nk} s_{0,n,2h_2} \left(J s_{n,2h_1} + m_{n,2h_1} \right) \end{aligned}$$

For Λ_n note that:

$$E[||b_i(\psi(\hat{\eta}) - \psi_i(\eta))||^2] \lesssim \xi_k^2 E[(\psi_i(\hat{\eta}) - \psi_i(\eta))^2].$$

Decomposing the second term on the right hand side exploiting Assumption A.3 together with independence of the nuisance models and the conditional independence of the potential outcomes yields:

$$\begin{aligned} E[(a.1)^2] &= E[(\hat{\mu}_i - \mu_i)^2 (1 - D_i/e_i)^2] \\ &= E[(\hat{\mu}_i - \mu_i)^2 (1 - 2 + 1/e_i)] \\ &\lesssim E[(\hat{\mu}_i - \mu_i)^2] \\ E[(a.2)^2] &= E[(\hat{e}_i - e_i)^2 (Y_i - \mu_i)^2 D_i / (e_i^2 \hat{e}_i^2)] \\ &= E[(\hat{e}_i - e_i)^2 E[(Y_i - \mu_i)^2 | X_i, D_i = 1] / (e_i \hat{e}_i^2)] \\ &\lesssim E[(\hat{e}_i - e_i)^2] \\ E[(a.3)^2] &= E[(\hat{\mu}_i - \mu_i)^2 (\hat{e}_i - e_i)^2 D_i / (e_i^2 \hat{e}_i^2)] \\ &= E[(\hat{\mu}_i - \mu_i)^2 (\hat{e}_i - e_i)^2 / (e_i \hat{e}_i^2)] \\ &\lesssim E[(\hat{\mu}_i - \mu_i)^2] \end{aligned}$$

where the last inequality uses a simple constant bound on the control propensities. Now note that here e_i denotes the aggregate treatment propensity with uniformly bounded inverse due to A.3. Thus, using the convergence rate for the aggregate mean $\hat{\mu}_i$ above, we

obtain that

$$\begin{aligned}\Lambda_n^{[nATE]} &= E[||b_i(\psi(\hat{\eta}) - \psi_i(\eta))||^2]^{1/2} \\ &\lesssim \xi_k\left(s_{0,n,2} + \sqrt{J}\left(\sup_{t\neq 0} s_{t,n,2} + \sup_{t\neq 0, x\in\mathcal{X}} m_{t,n,2} e_t(x)\right)\right) \\ &\lesssim \xi_k\left(s_{0,n,2} + \sqrt{J} s_{n,2} + J^{-1} m_{n,2}\right)\end{aligned}$$

by Assumption A.3.

A.2.3 Machine Learning Bias: $rATE$ and Δ

First note that, conditional on the event u_n

$$\begin{aligned}\max_{1\leq i\leq n} \frac{\pi_t}{\hat{e}_t(X_i)} &= \max_{1\leq i\leq n} \frac{e_t(X_i)}{\hat{e}_t(X_i)} \frac{\pi_t}{e_t(X_i)} \\ &\equiv \chi_{t,n} \\ &\lesssim_P \sup_{x\in\mathcal{X}} \sup_{\hat{e}_t\in E_{t,n}} \frac{e_t(x)}{\hat{e}_t(x)} \sup_{x\in\mathcal{X}} \frac{\pi_t}{e_t(x)} \\ &\lesssim_P \sup_{\hat{e}_t\in E_{t,n}, x\in\mathcal{X}} \frac{e_t(x)}{\hat{e}_t(x)} \\ &\lesssim_P 1\end{aligned}$$

by definition of the supremum and Assumptions A.3 and A.4. Note that this also implies that $\sup_{t\neq 0} \chi_{t,n} \lesssim_P 1$. Now, for the $rATE$, we exploit the same moment function decomposition as for the $nATE$ but for each potential outcome moment function within the weighted sum. Omitting the control part $\psi_i^{[0]}(\eta)$ again, we have that

$$\begin{aligned}\psi_i(\hat{\eta}, \pi) - \psi_i(\eta, \pi) = \left[\sum_{t\neq 0} \pi_t\right]^{-1} \sum_{t\neq 0} \pi_t \Bigg[&(\hat{\mu}_t(X_i) - \mu_t(X_i))\left(1 - \frac{D_{t,i}}{e_t(X_i)}\right) \\ &- (\hat{e}_t(X_i) - e_t(X_i))(Y_i - \mu_t(X_i)) \frac{D_{t,i}}{e_t(X_i)\hat{e}_t(X_i)} \\ &+ (\hat{\mu}_t(X_i) - \mu_t(X_i))(\hat{e}_t(X_i) - e_t(X_i)) \frac{D_{t,i}}{\hat{e}_t(X_i) e_t(X_i)}\Bigg] \\ = \left[\sum_{t\neq 0} \pi_t\right]^{-1} \sum_{t\neq 0} \pi_t &\left(A_{t,1} - A_{t,2} + A_{t,3}\right).\end{aligned}$$

We now consider $B_n^{[rATE]}$. Equivalently to the $nATE$ we have that

$$E[b_i A_{t,1}] = E[b_i A_{t,2}] = 0.$$

for all $t = 1, \ldots, J$. The remaining term can be bounded as

$$\begin{aligned}
&||E[b_i \sum_{t\neq 0} \pi_t A_{t,3}]|| \\
&\quad = \left|\left| E[b_i \sum_{t\neq 0} \pi_t (\hat{\mu}_t(X_i) - \mu_t(X_i))(\hat{e}_t(X_i) - e_t(X_i)) \frac{D_{t,i}}{\hat{e}_t(X_i) e_t(X_i)} \right]\right|| \\
&\quad \leq_P \sum_{t\neq 0} \pi_t \max_{1\leq i\leq n} e_t(X_i)^{-1} E[||b_i(\hat{\mu}_t(X_i) - \mu_t(X_i))(\hat{e}_t(X_i) - e_t(X_i))||] \\
&\quad \lesssim_P J \sup_{t\neq 0} \chi_{t,n} \sqrt{k} E[(\hat{\mu}_t(X_i) - \mu_t(X_i))^2 (\hat{e}_t(X_i) - e_t(X_i))^2]^{1/2} \\
&\quad \lesssim \sqrt{k} J s_{n,2h_1} m_{n,2h_2}
\end{aligned}$$

by Hölder's inequality with $1/h_1 + 1/h_2 = 1$. Thus, overall we have that

$$B_n^{[rATE]} \lesssim \sqrt{nk} J s_{n,2h_1} m_{n,2h_2}.$$

Now consider Λ_n. First note that

$$\begin{aligned}
&E[(\psi_i(\hat{\eta}, \pi) - \psi(\eta, \pi))^2] \\
&\quad = \left[\sum_{t\neq 0} \pi_t\right]^{-2} \sum_{t\neq 0} \sum_{s\neq 0} \pi_t \pi_s E\left[\left(A_{t,1} - A_{t,2} + A_{t,3}\right)\left(A_{s,1} - A_{s,2} + A_{s,3}\right)\right]
\end{aligned}$$

Now consider the summands for $s = t$:

$$\pi_t^2 E[(A_{t,1} - A_{t,2} + A_{t,3})^2] \lesssim 4\pi_t^2 E[A_{t,1}^2 + A_{t,2}^2 + A_{t,3}^2]$$

Bounding each term separately yields

$$\begin{aligned}
\pi_t^2 E[A_{t,1}^2] &= \pi_t^2 E[(\hat{\mu}_t(X_i) - \mu_t(X_i))^2(1 - 2 + 1/e_t(X_i))] \\
&\lesssim \pi_t \left(\sup_{x \in \mathcal{X}} \left| \frac{\pi_t}{e_t(x)} \right| + \pi_t \right) E[(\hat{\mu}_t(X_i) - \mu_t(X_i))^2] \\
&\lesssim J^{-1} m_{t,n,2}^2 \\
\pi_t^2 E[A_{t,2}^2] &= \pi_t^2 E[(\hat{e}_t(X_i) - e_t(X_i))^2 E[(Y_i(t) - \mu_t(X_i))^2 | X_i]/(e_t(X_i)\hat{e}_t(X_i)^2)] \\
&\lesssim_P \pi_t^{-1} \sup_{x \in \mathcal{X}} \frac{\pi_t^3}{e_t(x)^3} \chi_{t,n}^2 E[(\hat{e}_t(X_i) - e_t(X_i))^2] \\
&\leq J s_{t,n,2}^2 \\
\pi_t^2 E[A_{t,3}^2] &= \pi_t^2 E[(\hat{e}_t(X_i) - e_t(X_i))^2 (\hat{\mu}_t(X_i) - \mu_t(X_i))^2/(e_t(X_i)\hat{e}_t(X_i)^2)] \\
&\leq J s_{t,n,2}^2
\end{aligned}$$

by Assumptions A.3 and A.4. Now consider the summands with $s \neq t$. Recall that $D_{t,i} D_{s,i} = 0$. As a preliminary, note that (conditional on the cross-fitted model):

$$\begin{aligned}
E\left[\left(1 - \frac{D_{t,i}}{e_t(X_i)} \right) \left(1 - \frac{D_{s,i}}{e_s(X_i)} \right) \middle| X_i \right] &= -1 \\
E\left[\left(1 - \frac{D_{t,i}}{e_t(X_i)} \right) \frac{D_{s,i}}{e_s(X_i)\hat{e}_s(X_i)} \middle| X_i \right] &= \frac{1}{\hat{e}_s(X_i)} \\
E\left[\left(1 - \frac{D_{t,i}}{e_t(X_i)} \right) D_{s,i}(Y_i(s) - \mu_s(X_i)) \middle| X_i \right] &= 0
\end{aligned}$$

Thus, summands simplify to

$$\begin{aligned}
&\pi_t \pi_s E\Bigg[- (\hat{\mu}_t(X_i) - \mu_t(X_i))(\hat{\mu}_s(X_i) - \mu_s(X_i)) \\
&\quad + (\hat{\mu}_t(X_i) - \mu_t(X_i))(\hat{\mu}_s(X_i) - \mu_s(X_i)) \left(\frac{\hat{e}_t(X_i) - e_t(X_i)}{\hat{e}_t(X_i)} + \frac{\hat{e}_s(X_i) - e_s(X_i)}{\hat{e}_s(X_i)} \right) \Bigg] \\
&\quad \lesssim \left[\pi_t \pi_s m_{t,n,2} m_{s,n,2} \right] + \left[\left(\pi_s s_{t,n,2} \chi_{t,n} + \pi_t s_{s,n,2} \chi_{s,n} \right) m_{t,n,2h_1} m_{s,n,2h_2} \right]
\end{aligned}$$

by repeated application of Hölder's inequality with $1/h_1 + 1/h_2 = 1$. Note that we have J variances and $J(J-1)$ times the covariance term. Thus, we obtain

$$\begin{aligned}\Lambda_n^{[rATE]} &\lesssim \xi_k E[(\psi_i(\hat{\eta}, \pi) - \psi(\eta, \pi))^2]^{1/2} \\ &\lesssim_p \xi_k \left(m_{n,2}^2 + J^2 s_{n,2}^2 + J(J-1)(J^{-2} m_{n,2}^2) + J(J-1)(J^{-1} s_{n,2} m_{n,2h_1} m_{n,2h_2}) \right)^{1/2} \\ &\lesssim \xi_k (m_{n,2} + J s_{n,2} + \sqrt{J s_{n,2} m_{n,2h_1} m_{n,2h_2}})\end{aligned}$$

A.3 Asymptotic Linearization and Normality

Without loss of generality, we set $Q = I$ but assume a random design, i.e. Q is unknown as in BCCK. Derivations are split by $nATE$ and $rATE/\Delta$ and thus quantities depending on $\psi_i(\cdot)$, r_i, ε_i, l_k, c_k, β_0 are decomposition parameter specific if not stated otherwise.

A.3.1 Linearization

We first expand the $\hat{Q}$-weighted estimator around the best linear predictor. Then, we control for the machine learning bias and the additional uncertainty from estimating π. In the end, we combine these rates with the rates from estimating $\hat{Q}$ to obtain an asymptotic linearization of the series estimator.

$\underline{rATE/\Delta}$**:** Take a sequence of basis function $b = b_n$ such that $||b|| = 1$. Decompose

$$\begin{aligned}\sqrt{n} b'(E_n[b_i \psi_i(\hat{\eta}, \hat{\pi})] - \hat{Q}\beta_0) = &\sqrt{n} b' E_n[b_i(\psi_i(\eta, \pi) - b_i'\beta_0)] \\ &+ \sqrt{n} b' E_n[b_i(\psi_i(\hat{\eta}, \pi) - \psi_i(\eta, \pi))] \\ &+ \sqrt{n} b' E_n[b_i(\psi_i(\hat{\eta}, \hat{\pi}) - \psi_i(\hat{\eta}, \pi))]\end{aligned}$$

The first term will be part of the first order asymptotics used for the normality results later. Now by the derivation in Section A.2.3 and Chebyshev's inequality we have that

$$||\sqrt{n} b' E_n[b_i(\psi_i(\hat{\eta}, \pi) - \psi_i(\eta, \pi))]|| \lesssim_P B_n^{[rATE]} + \Lambda_n^{[rATE]}$$

The third term can further be decomposed exploiting the multiplicative structure and (H.6):

$$\begin{aligned}
&\sqrt{n}E_n[b_i(\psi_i(\hat{\eta},\hat{\pi})-\psi_i(\hat{\eta},\pi))] \\
&\quad=\sqrt{n}E_n\left[b_i\sum_{t\neq 0}\psi_i^{[t]}(\hat{\eta})\left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0}-\frac{\pi_t}{1-\pi_0}\right)\right] \\
&\quad=\sum_{t\neq 0}E_n[b_i\psi_i^{[t]}(\hat{\eta})]\sqrt{n}\left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0}-\frac{\pi_t}{1-\pi_0}\right) \\
&\quad=\sum_{t\neq 0}(\hat{\gamma}_t-\gamma_t+\gamma_t)\left(\frac{1-\pi_0}{1-\hat{\pi}_0}-1+1\right)G_n[a_i^{[t]}] \\
&\quad=(\hat{\gamma}-\gamma)G_n[a_i]\left(\frac{1-\pi_0}{1-\hat{\pi}_0}-1+1\right)+\left(\frac{1-\pi_0}{1-\hat{\pi}_0}-1\right)\gamma G_n[a_i]+\gamma G_n[a_i]
\end{aligned}$$

where $\gamma_t = E[b_i\psi_i^{[t,0]}(\eta)] = E[b_i\tau_t(X_i)]$, $\gamma = (\gamma_1 \dots \gamma_J)$, and $a_i = (a_i^{[1]} \dots a_i^{[J]})'$. Now note that, by the iid assumption, Chebyshev's inequality yields

$$||G_n[a_i]|| \lesssim_P J \sup_{t\neq 0} E[(a_i^{[t]})^2] \lesssim J \sup_{t\neq 0}\pi_t \lesssim 1$$

Thus, by (H.5), we obtain that

$$\sqrt{n}b'E_n[b_i(\psi_i(\hat{\eta},\hat{\pi})-\psi_i(\hat{\eta},\pi))] = b'\gamma G_n[a_i] + R_{n,\pi}$$

where

$$\begin{aligned}
||R_{n,\pi}|| &\lesssim_P ||\hat{\gamma}-\gamma||\ ||G_n[a_i]||\ (1+n^{-1/2}) + n^{-1/2}||\gamma||\ ||G_n[a_i]|| \\
&\lesssim_P J\sqrt{k}(n^{-1/2}+m_{n,2}+Js_{n,2}) + n^{-1/2}\sqrt{Jk} \\
&\lesssim J\sqrt{k}(n^{-1/2}+m_{n,2}+Js_{n,2})
\end{aligned}$$

as $||\frac{1-\pi_0}{1-\hat{\pi}_0}-1|| \lesssim_P n^{-1/2}$ as $1-\pi_0$ is bounded away from zero. Recall that $\psi_i(\eta,\pi) - b_i'\beta_0 = \varepsilon_i + r_i$. Now what is left is the remainder due to estimation using $\hat{Q}$. We make use of

(H.4) for bounding $||\hat{Q} - I||$. Conditional on the data, observe that

$$\begin{aligned} V[b'(\hat{Q}^{-1} - I)G_n[b_i\varepsilon_i]|Z_1, \ldots, Z_n] &\lesssim b'(\hat{Q}^{-1} - I)\hat{Q}(\hat{Q}^{-1} - I)b \\ &\lesssim_P \frac{\xi_k^2 \log k}{n} \end{aligned}$$

Moreover,

$$|b'(\hat{Q}^{-1} - I)G_n[b_i r_i]| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} l_k c_k \sqrt{k}$$

as in BCCK, Proof of Lemma 4.1 and using (H.7). For the final term note that $\sup_{z \in \mathcal{Z}} ||E[a_i a_i'|Z_i = z]|| \lesssim J^{-1/2}$ as in (H.5). This yields

$$\begin{aligned} V[b'(\hat{Q}^{-1} - I)G_n[\gamma a_i]|Z_1, \ldots, Z_n] &= b'(\hat{Q}^{-1} - I)\gamma E[a_i a_i'|Z_i]\gamma'(\hat{Q}^{-1} - I)b \\ &\lesssim_P ||\hat{Q}^{-1}||^2 \; ||\hat{Q}^{-1} - I||^2 || \; ||\gamma||^2 J^{-1/2} \\ &\lesssim \frac{\xi_k^2 \log k}{n} k\sqrt{J} \end{aligned}$$

Thus, decomposing and centering the linear predictor yields:

$$\sqrt{n}b'(\hat{\beta} - \beta_0) = b'G_n[b_i(\varepsilon_i + r_i) + \gamma a_i] + R_{n,Q} + R_{n,\pi} + R_{n,\eta}$$

where

$$\begin{aligned} ||R_{n,Q}|| &\lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}} \left(1 + k^{1/2}(J^{1/4} + l_k c_k)\right) \\ ||R_{n,\pi}|| &\lesssim_P J\sqrt{k}(n^{-1/2} + m_{n,2} + J s_{n,2}) \\ ||R_{n,\eta}|| &\lesssim_P B_n^{[rATE]} + \Lambda_n^{[rATE]} \end{aligned}$$

<u>*nATE*</u>**:** For the *nATE* there is no $G_n[a_i]$ term as there are no unconditional probability terms π to be estimated. Moreover, the machine-learning approximation rates are different as shown above. All other derivations follow along the same lines. Thus, we obtain for the

nATE

$$\sqrt{n}b'(\hat{\beta} - \beta_0) = b'G_n[b_i(\varepsilon_i + r_i)] + R_{n,Q} + R_{n,\eta}$$

where

$$||R_{n,Q}|| \lesssim_P \sqrt{\frac{\xi_k^2 \log k}{n}}\left(1 + k^{1/2} l_k c_k\right)$$
$$||R_{n,\eta}|| \lesssim_P B_n^{[nATE]} + \Lambda_n^{[nATE]}$$

where here $\varepsilon_i = \psi_i^{[nATE]}(\eta) - E[\psi_i^{[nATE]}(\eta)|Z_i]$ and $r_i = E[\psi_i^{[nATE]}(\eta)|Z_i] - b_i'\beta_0$.

A.3.2 Asymptotic Normality

rATE/Δ: Again let $Q = I$ without loss of generality. Under Assumption C.4 all remainders from the previous subsection are $o_p(1)$ and we are left with leading term

$$\frac{b'G_n[b_i(\varepsilon_i + r_i) + \gamma a_i]}{||b'\Omega^{1/2}||} = \sum_{i=1}^{n} \frac{b'[b_i(\varepsilon_i + r_i) + \gamma a_i]}{\sqrt{n}||b'\Omega^{1/2}||}$$

We now verify the Lindeberg condition for asymptotic normality. First note that the term above has expectation zero. By independence and the binomial formula we obtain for any $\delta > 0$

$$\begin{aligned}
&\sum_i^n E\left[\frac{(b'(b_i(\varepsilon_i + r_i) + \gamma a_i))^2}{nb'\Omega b}\mathbb{1}\left(\frac{|b'(b_i(\varepsilon_i + r_i) + \gamma a_i)|}{\sqrt{n}||b'\Omega^{1/2}||} > \delta\right)\right] \\
&\quad \leq 4nE\left[\frac{(b'b_i(\varepsilon_i + r_i))^2}{nb'\Omega b}\mathbb{1}\left(\frac{|b'b_i(\varepsilon_i + r_i)|}{\sqrt{n}||b'\Omega^{1/2}||} > \frac{\delta}{2}\right)\right] + 4nE\left[\frac{(b'\gamma a_i)^2}{nb'\Omega b}\mathbb{1}\left(\frac{|b'\gamma a_i|}{\sqrt{n}||b'\Omega^{1/2}||} > \frac{\delta}{2}\right)\right] \\
&\quad \equiv (an.1) + (an.2)
\end{aligned}$$

Now denote

$$w_{ni} := \frac{b'b_i}{||b'\Omega_1^{1/2}||} \;\Rightarrow\; |w_{ni}| \lesssim \frac{\xi_k}{\sqrt{n}}, \quad nE[|w_{ni}|^2] \lesssim 1$$

analogously to BCCK Proof of Theorem 4.2 using the conditional moment bound (H.7). Now note that, by the eigenvalue assumption C.6, (an.1) is bounded by:

$$
\begin{aligned}
(an.1) &\lesssim \frac{nb'\Omega_1 b}{nb'\Omega b} E\left[\frac{(b'b_i(\varepsilon_i + r_i))^2}{nb'\Omega_1 b}\mathbb{1}\left(\frac{|b'b_i(\varepsilon_i + r_i)|}{\sqrt{n}||b'\Omega_1^{1/2}||} > \frac{\delta}{2}\frac{||b'\Omega^{1/2}||}{||b'\Omega_1^{1/2}||}\right)\right] \\
&\lesssim 2nE[|w_{ni}|^2\varepsilon_i^2\mathbb{1}(|\varepsilon_i| + |r_i| > \delta/|w_{ni}|)] + 2nE[|w_{ni}|^2 \sup_{z\in\mathcal{Z}}|r(z)|^2\mathbb{1}(|\varepsilon_i| + |r_i| > \delta/|w_{ni}|)] \\
&\equiv (an.1.i) + (an.1.ii)
\end{aligned}
$$

Using C.2 $\sup_z |r(z)| \leq l_k c_k$ and (H.9), we obtain that

$$
\begin{aligned}
(an.1.i) &\lesssim nE[|w_{ni}|^2 E[\varepsilon_i^2\mathbb{1}(|\varepsilon_i| > (\delta\sqrt{n}/(2c\xi_k) - l_k c_k))|Z_i]] \\
&\lesssim \sup_{z\in\mathcal{Z}} E\left[\sup_{t\neq 0}\frac{\varepsilon_i(t)^2 D_{t,i}}{e_t(X_i)^2}\mathbb{1}\left(\frac{|\varepsilon_i(t)|D_{t,i}}{e_t(X_i)} > \frac{\delta\sqrt{n}}{2\xi_k} - l_k c_k\right)\middle| Z_i = z\right] \\
&\lesssim \sup_{z\in\mathcal{Z}}\sup_{t\neq 0}\sup_{x\in\mathcal{X}}\frac{\pi_t^2}{e_t(x)^2}\pi_t^{-2}E[\sup_{t\neq 0}\varepsilon_i(t)^2 D_{t,i}\mathbb{1}(|\varepsilon_i(t)|D_{t,i} > c_n/J)|Z_i = z]
\end{aligned}
$$

where $c_n = \left(\frac{\delta\sqrt{n}}{2\xi_k} - l_k c_k\right)$. Now by the integrated tail formula we have that

$$
\begin{aligned}
&E[\sup_{t\neq 0}\ \varepsilon_i(t)^2 D_{t,i}\mathbb{1}(|\varepsilon_i(t)|D_{t,i} > c_n/J|Z_i = z)] \\
&= \int_0^\infty P(\sup_{t\neq 0}\varepsilon_i(t)^2 D_{t,i}\mathbb{1}(|\varepsilon_i(t)|D_{t,i} > c_n/J) > w|Z_i = z)dw \\
&\leq \sum_{t\neq 0} P(D_{t,i} = 1|Z_i = z)\int_0^\infty P(\varepsilon_i(t)^2\mathbb{1}(|\varepsilon_i(t)| > c_n/J) > w|Z_i = z, D_{t,i} = 1)dw \\
&\lesssim J\sup_{t\neq 0}\pi_t\int_0^\infty P(\varepsilon_i(t)^2\mathbb{1}(|\varepsilon_i(t)| > c_n/J) > w|Z_i = z, D_{t,i} = 1)dw \\
&\lesssim \sup_{t\neq 0} E[\varepsilon_i(t)^2\mathbb{1}(|\varepsilon_i(t)| > c_n/J)|Z_i = z, D_{t,i} = 1]
\end{aligned}
$$

Note that, by Markov's inequality and the conditional m-th moment bound in C.1, we have that

$$
P(|\varepsilon_i(t)| > c_n/J|Z_i = z, D_{t,i} = 1) \lesssim (c_n/J)^{-m}
$$

Hölder's inequality then yields for the equation above

$$E[\varepsilon_i(t)^2 \mathbb{1}(|\varepsilon_i(t)| > c_n/J)|Z_i = z, D_{t,i} = 1] \leq \left(E[|\varepsilon_i(t)|^m | Z_i = z, D_{t,i} = 1](c_n/J)^{-m} \right)^{2/m}$$
$$\lesssim \left[\left(\frac{\delta\sqrt{n}}{2c\xi_k} - l_k c_k \right) \frac{1}{J} \right]^{-2}$$

for any $t \neq 0$. Plugging this back into $(an.1.i)$ and using A.3 $\sup_{t\neq 0} \pi_t J \lesssim 1$ yields

$$(an.1.i) \lesssim \left[\left(\frac{\delta\sqrt{n}}{2c\xi_k} - l_k c_k \right) \frac{1}{J} \right]^{-2} J^2$$
$$\lesssim \frac{\xi_k^2 J^4}{n}$$

As $\delta\sqrt{n}/\xi_k - l_k c_k \to \infty$. For $(an.1.ii)$ it follows equivalently to BCCK, Proof of Theorem 4.2, that

$$(an.1.ii) \lesssim l_k^2 c_k^2 \sup_{z\in\mathcal{Z}} P(|\varepsilon_i| > c\delta\sqrt{n}/\xi_k - l_k c_k | Z_i = z)$$

Analogously to the derivations for $(an.1.i)$ using (H.9) we obtain that

$$P(|\varepsilon_i| > c_n | Z_i = z) \lesssim \sum_{t\neq 0} P(|\varepsilon_i(t)| > c_n/J | Z_i = z, D_{t,i} = 1) P(D_{t,i} = 1 | Z_i = z)$$
$$\lesssim \sup_{t\neq 0} P(|\varepsilon_i(t)| > c_n/J | Z_i = z, D_{t,i} = 1)$$
$$\lesssim \sup_{t\neq 0} \frac{E[|\varepsilon_i(t)|^m | Z_i = z, D_{t,i} = 1]}{(c_n/J)^m}$$
$$\lesssim (c_n/J)^{-m}$$

where the last two steps follow from Markov's inequality and the conditional moment bound. Plugging this back into $(an.1.ii)$ yields

$$(an.1.ii) \lesssim \left(\frac{(l_k c_k)^{\frac{2}{m}} J}{[\delta\sqrt{n}/\xi_k - l_k c_k]} \right)^m$$

Now consider $(an.2)$. Note that by (H.5), we have that

$$\begin{aligned}
(an.2) &\leq \frac{||\gamma||^2}{b'\Omega b} E[||a_i||^2 \mathbb{1}(||a_i|| > (\delta/2)||b'\Omega^{1/2}||/||\gamma||)] \\
&\lesssim ||\gamma||^2 E[||a_i||^2 \mathbb{1}(||a_i|| > C\delta\sqrt{n/kJ})] \\
&\lesssim kJ \sum_{t\neq 0} E[(a_i^{[t]})^2 \mathbb{1}(||a_i|| > C\delta\sqrt{n/kJ})] \\
&\lesssim kJ^2 P(||a_i|| > C\delta\sqrt{n/kJ}) \\
&\lesssim kJ^2 P(\sup_{t\neq 0}(a_i^{[t]})\sqrt{J} > C\delta\sqrt{n/kJ}) \\
&\lesssim kJ^3 \sup_{t\neq 0} P(|a_i^{[t]}| > C\delta\sqrt{n/kJ^2}) \\
&= o(1)
\end{aligned}$$

if $\sqrt{n/kJ^2} \to \infty$ as $a_i^{[t]}$ is uniformly bounded for all t. Thus using Assumption C.5 yields

$$(an.1) + (an.2) \lesssim \frac{\xi_k^2 J^4}{n} + \left(\frac{(l_k c_k)^{\frac{2}{m}} J}{[\delta\sqrt{n}/\xi_k - l_k c_k]}\right)^m = o(1)$$

<u>$nATE$</u>**:** First note that there is no $(an.2)$ term for the $nATE$. For $(an.1)$ most derivations follow analogously. However, due to the propensity score weighting we can use (H.9) and improve on some of the rates in $(an.1.i)$

$$\begin{aligned}
E[\varepsilon_i^2 \mathbb{1}(|\varepsilon_i| > c_n)|Z_i = z] &\lesssim E[\sup_{t\neq 0} \varepsilon_i(t)^2 \mathbb{1}(|\varepsilon_i(t)| > c_n)|Z_i = z] \\
&\lesssim_P J \sup_{z\in\mathcal{Z}} \sup_{t\neq 0} E[\varepsilon_i(t)^2 \mathbb{1}(|\varepsilon_i(t)| > c_n)|Z_i = z] \\
&\lesssim J c_n^{-2}
\end{aligned}$$

Similarly for $(an.1.ii)$, we have that

$$\begin{aligned}
P(|\varepsilon_i| > c_n) &\leq J \sup_{t\neq 0} P(|\varepsilon_i(t)| > c_n) \\
&\lesssim J c_n^{-m}
\end{aligned}$$

Plugging both into $(an.1)$ then with c_n as above yields

$$(an.1) \lesssim J\frac{}{[\delta\sqrt{n}/\xi_k - l_k c_k]^2} + J\frac{l_k^2 c_k^2}{[\delta\sqrt{n}/\xi_k - l_k c_k]^m} = o(1)$$

under the assumption B.5. Note that convergence is faster than $(an.1)$ for the $rATE$. Asymptotic normality then follows from the sufficiency of the Lindeberg condition. □

B Supplementary Appendix

B.1 Toy example

Consider a setting with a binary heterogeneity variable $X_i \in \{0, 1\}$ and three effective treatments $T_i \in \{0, 1, 2\}$. We impose deterministic potential outcomes that are homogeneous within treatment status, but heterogeneous between treatments:

	$Y_i(0)$	$Y_i(1)$	$Y_i(2)$
$X_i = 0$	0	-1	1
$X_i = 1$	0	-1	1

Both groups defined by X_i have the same potential outcomes under the different treatments. This means there can be no real effect heterogeneity. However, consider now that the probability to receive the effective treatments varies with X_i:

	$P(T_i = 0\|X_i)$	$P(T_i = 1\|X_i)$	$P(T_i = 2\|X_i)$
$X_i = 0$	0.5	1/8	3/8
$X_i = 1$	0.5	3/8	1/8

Collapsing treatments one and two into a binary treatment $D_i = \mathbb{1}(T_i > 0)$ and running a subgroup analysis for the "treatment" D_i results in the following conditional average treatment effects ($CATE$):

$$CATE(X_i) = 1 - 4 \cdot P(T_i = 1|X_i) = \begin{cases} 0.5 & \text{if } X_i = 0 \\ -0.5 & \text{if } X_i = 1. \end{cases}$$

Thus, the aggregation into the binary indicator leads us to "find" a positive effect for one group and a negative effect for another group although the effective treatments actually do not create heterogeneous effects. Everything is just driven by them receiving a different mix of effective treatments.

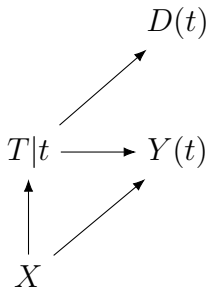
B.2 Identification

B.2.1 Conditional independencies

We can read off the conditional independencies with respect to potential, not observed, outcomes encoded in DAGs (1) and (2) from single-world intervention graphs (SWIG) of Richardson and Robins (2013). We intervene on T_i to read off the independencies we require for identification of our decompositions.

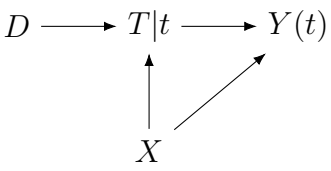
Scenario 1:

Figure B.1: SWIG with intervention on T_i



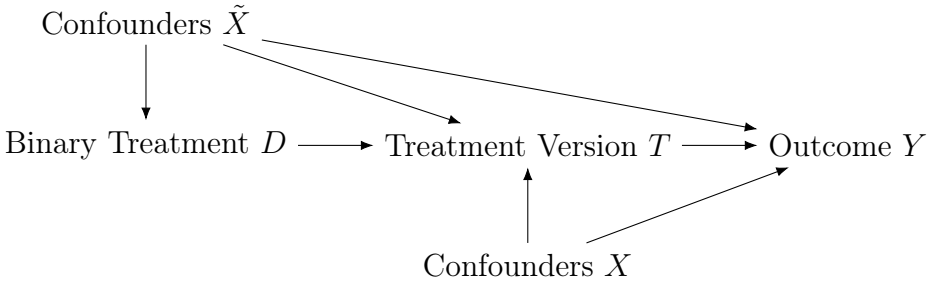
Scenario 2:

Figure B.2: SWIG with intervention on T_i



Both SWIGs (B.1) and (B.2) imply the conditional independence shown in Equation (2). This links the observed effective treatment to the unobserved potential outcomes and justifies our assumption 1a required for identifying our decomposition terms.

Figure B.3: DAG: Confounded binary treatment precedes confounded treatment version:



B.2.2 Identification in Scenario 2 with confounded binary treatment

Figure B.3 considers the case of a binary treatment that is potentially confounded with both treatment version selection and outcome due to the backdoor through $\tilde{X}$. This could occur e.g. in the case of the evaluation of Job Corps access on earnings when access was not allocated randomly but is based on observables. In this case, the derived conditional independence assumptions change to

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp D_i \mid T_i, \tilde{X}_i \tag{14}$$

$$Y_i(0), Y_i(1) \ldots, Y_i(J) \perp\!\!\!\perp T_i \mid X_i, \tilde{X}_i \tag{15}$$

Thus, the adjustment set required for identification and entering estimation of the nuisance parameters of the decomposition terms needs to incorporate the additional set of confounders $\tilde{X}_i$, but all results hold equivalently.

B.3 Connection to Continuous Treatments

The unconditional $rATE$ parameter can be seen as a discrete approximation to the integrated continuous effect curve under the normalization that $\mu_0 = 0$. In particular, assume that the treatment now is continuous, i.e. $t \in J^* \subset \mathbb{R}$, with J^* being a compact subset of the real line, e.g. $[0, 1]$. Kennedy et al. (2017), Equation (2) denotes the integrated effect curve of a continuous treatment as

$$\int_{J^*} \int_{\mathcal{X}} \mu(t, x) \pi(t) dF(x) dt = \int_{J^*} \mu(t) dt \tag{16}$$

where $\mu(t,x) = E[Y_i(t)|X_i = x]$ and $\pi(t)$ being the marginal treatment density. Now consider a discretization of the support J^* using J steps that define J multi-valued treatments, e.g $\{1/J, 2/J, \dots, 1\}$. Let μ_t and π_t be the corresponding discrete multi-valued potential outcomes and cell probability/ propensity score. If $\mu(t)$ is continuous, then it is straightforward to show that

$$\lim_{J\to\infty} rATE = \lim_{J\to\infty} \frac{\sum_{j=1}^{J} \pi_t \mu_t}{\sum_{j=1}^{J} \pi_t} = \int_{J^*} \mu(t)dt \tag{17}$$

Thus, when we allow for $J \to \infty$, the $rATE$ can arbitrarily well approximate/nest the integrated continuous effect curve. The explicit choice of discretization corresponds to the bandwidth selection that has to be made when estimating this quantity (Kennedy et al., 2017).

B.4 Neyman-orthogonality

The key insight required here is that the $nATE$ and $rATE$ scores are Neyman-orthogonal with known unconditional probabilities π_t, $t = 1, \dots, J$. We show how the additional estimation error can be incorporated in Appendix A. Here we are concerned with the Gateaux derivative of the $nATE$ and $rATE$ scores with respect to the vector of infinite-dimensional nuisance parameters $\eta = (\mu(x), p(x)) = (\mu_0(x), \dots, \mu_J(x), e_0(x), \dots, e_J(x))'$. As π is assumed to be known, we suppress dependence $\psi(\eta, \pi) = \psi(\eta)$ out of convenience for now. Suppressing also the dependencies of the nuisance parameters on x, we write the path-wise derivative of the conditional expectation of a score with respect to the vector of nuisance parameters as

$$\partial_\eta E[\psi_i(\eta)|X_i = x] = \partial_r E[\psi_i(\dots, \mu_t + r(\tilde{\mu}_t - \mu_t), \dots, e_t + r(\tilde{e}_t - e_t), \dots)|X_i = x]|_{r=0}$$

First, we revisit Neyman-orthogonality of the doubly robust score:

$$
\begin{aligned}
&\partial_r E[\psi_i^{[t]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \partial_r E\left[(\mu_t + r(\tilde{\mu}_t - \mu_t)) + \frac{D_{t,i}Y_i}{e_t + r(\tilde{e}_t - e_t)} - \frac{D_{t,i}(\mu_t + r(\tilde{\mu}_t - \mu_t))}{e_t + r(\tilde{e}_t - e_t)}\middle| X_i = x\right]\Bigg|_{r=0} \\
&\quad = (\tilde{\mu}_t - \mu_t) - \frac{e_t\mu_t(\tilde{e}_t - e_t)}{e_t^2} - \frac{e_t^2(\tilde{\mu}_t - \mu_t) - e_t\mu_t(\tilde{e}_t - e_t)}{e_t^2} \\
&\quad = 0
\end{aligned}
$$

where we use that $E[D_{t,i}Y_i|X_i = x] = E[D_{t,i}\sum_t D_{t,i}Y_i(t)|X_i = x] = E[D_{t,i}Y_i(t)|X_i = x] = e_t\mu_t$ by the observational rule and Assumption 1.

B.4.1 rATE

As the $rATE$ score is a linear combination of doubly robust scores, it inherits the Neyman-orthogonality of its components:

$$
\begin{aligned}
&\partial_r E[\psi_i^{[rATE]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \sum_{t\neq 0} \frac{\pi_t}{1 - \pi_0}\partial_r E[\psi_i^{[t]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\qquad - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = 0
\end{aligned}
$$

B.4.2 nATE

The $nATE$ score differs from the standard doubly robust scores but can still be shown to be Neyman-orthogonal:

$$
\begin{aligned}
&\partial_r E[\psi_i^{[nATE]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \partial_r E\left[\frac{\sum_{t\neq 0}[(\mu_t + r(\tilde{\mu}_t - \mu_t))(e_t + r(\tilde{e}_t - e_t))]}{\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))} + \frac{D_i Y_i}{\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))}\right. \\
&\qquad \left. - \frac{D_i \sum_{t\neq 0}[(\mu_t + r(\tilde{\mu}_t - \mu_t))(e_t + r(\tilde{e}_t - e_t))]}{[\sum_{t\neq 0}(e_t + r(\tilde{e}_t - e_t))]^2}\middle| X_i = x\right]\Bigg|_{r=0} \\
&\quad - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = \frac{\sum_{t\neq 0}[\mu_t(\tilde{e}_t - e_t) + e_t(\tilde{\mu}_t - \mu_t)]\sum_{t\neq 0} e_t}{[\sum_{t\neq 0} e_t]^2} - \frac{\sum_{t\neq 0}\mu_t e_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} \\
&\quad - \frac{\sum_{t\neq 0} e_t\mu_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} - \frac{\sum_{t\neq 0} e_t \sum_{t\neq 0}[\mu_t(\tilde{e}_t - e_t) + e_t(\tilde{\mu}_t - \mu_t)]}{[\sum_{t\neq 0} e_t]^2} \\
&\quad + 2\frac{\sum_{t\neq 0}\mu_t e_t \sum_{t\neq 0}(\tilde{e}_t - e_t)}{[\sum_{t\neq 0} e_t]^2} - \partial_r E[\psi_i^{[0]}(\eta + r(\tilde{\eta} - \eta))|X_i = x]|_{r=0} \\
&\quad = 0
\end{aligned}
$$

where we use that $E[D_i Y_i|X_i = x] = E[D_i \sum_t D_{t,i} Y_i(t)|X_i = x] = \sum_{t\neq 0} e_t\mu_t$ by the observational rule and Assumption 1. Consequently, the difference between the $nATE$ and $rATE$ score that forms the Δ score is Neyman-orthogonal as well:

$$\partial_\eta E[\psi_i^{[nATE]}(\eta) - \psi_i^{[rATE]}(\eta)|X_i = x] = 0$$

B.5 Estimation of Asymptotic Variance

Let $E_n[X_i] = \frac{1}{n}\sum_{i=1}^n X_i$. Define

$$
\begin{aligned}
\hat{Q} &= E_n[b(Z_i)b(Z_i)'] \\
\hat{\Omega} &= \hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}
\end{aligned}
\tag{18}
$$

For the $rATE$ we use

$$\hat{\Sigma} = E_n\left[(b(Z_i)e_i + \hat{a}_i - \bar{\hat{a}}_i)(b(Z_i)e_i + \hat{a}_i - \bar{\hat{a}}_i)'\right]$$

with

$$\begin{aligned}
e_i &= \psi_i^{[rATE]}(\hat{\eta}, \hat{\pi}) - b(Z_i)'\hat{\beta} \\
\hat{a}_i &= \sum_{t\neq 0} \frac{E_n[b(Z_i)(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[0]}(\hat{\eta}))](D_{t,i}(1-\hat{\pi}_0) + D_{0,i}\hat{\pi}_t)}{(1-\hat{\pi}_0)^2} \\
\bar{\hat{a}}_i &= E_n[\hat{a}_i] \\
\hat{\pi}_t &= E_n[D_{t,i}]
\end{aligned}$$

For Δ, the $\psi_i^{[rATE]}(\hat{\eta}, \hat{\pi})$ has to be replaced by the corresponding score function and $\hat{\Sigma}$ changes to

$$\hat{\Sigma} = E_n\left[(b(Z_i)e_i - \hat{a}_i + \bar{\hat{a}}_i)(b(Z_i)e_i - \hat{a}_i + \bar{\hat{a}}_i)'\right].$$

For the $nATE$ we use only $\hat{\Sigma} = E_n[b(Z_i)b(Z_i)'e_i^2]$ as there are no estimated unconditional weights.

B.6 Asymptotic Variance Estimation Theory

In the following, we use some of the Lemmas in BCCK. To do so, we impose the following additional assumption. We require that $\xi_k^{2m/(m-2)} \log k/n \lesssim 1$, $\log \xi_k \lesssim \log k$ and Lipschitz constant

$$\xi_k^L := \sup_{x,x'\in\mathcal{X}, x\neq x'} \frac{||b(x) - b(x')||}{||x - x'||}$$

obeys Condition (A.5) from BCCK, i.e. $\log \xi_k^L \lesssim \log k$. Moreover, $c_k l_k \lesssim \sqrt{\log k}$ and $\sqrt{\frac{\xi_k^2 \log k}{n}}\left((nJ)^{1/m}\sqrt{\log k} + \sqrt{k} l_k c_k\right) \lesssim \sqrt{\log k}$ as in Theorem 4.6 by BCCK. Moreover, we assume $M_{n,1} = o(1)$ for the $nATE$ or $M_{n,2} = o(1)$ for $rATE/\Delta$ respectively where $M_{n,1}$ and $M_{n,2}$ are defined at the end of the section in equations (20) and (19) respectively. We

now provide first some auxiliary results and then the derivations for $rATE/\Delta$. The rates for $nATE$ follow directly by simplification.

B.6.1 Auxiliary Results

(MA.1) $\psi_i^{[t]}$ **bounds** First note that, for any $t \neq 0$,

$$\begin{aligned}|\psi_i^{[t]}(\eta)| &= \left|\frac{D_{t,i}\varepsilon_i(t)}{e_t(X_i)} + \mu_t(X_i)\right| \\ &\leq \sup_{x\in\mathcal{X}} \frac{\pi_t}{e_t(x)}\pi_t^{-1}|D_{t,i}||\varepsilon_i(t)| + |\mu_t(X_i)|\end{aligned}$$

Thus

$$\begin{aligned}\max_{1\leq i\leq n} |\psi_i^{[t]}(\eta)| &\lesssim_P \sup_{x\in\mathcal{X}} \frac{\pi_t}{e_t(x)}\pi_t^{-1} \max_{1\leq i\leq n} |\varepsilon_i(t)| + \sup_{x\in\mathcal{X}} |\mu_t(x)| \\ &\lesssim_P Jn^{1/m}\end{aligned}$$

by the Assumption A.2 and B.1/C.1. Moreover

$$\begin{aligned}|\psi_i^{[t]}(\eta)\psi_i^{[t']}(\eta)| &\leq |\psi_i^{[t]}(\eta)|^2 + |\psi_i^{[t']}(\eta)|^2 \\ &\leq 2\sup_{t\neq 0} |\psi_i^{[t]}(\eta)|^2\end{aligned}$$

and similarly

$$|\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta)||\psi_i^{[t']}(\hat{\eta}) - \psi_i^{[t']}(\eta)| \leq 2\sup_{t\neq 0} |\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta)|^2$$

Now consider the product of the moment functions for any η. For the square note that

$$\begin{aligned}|\psi_i^{[t]}(\eta)|^2 &= \left|\frac{D_{t,i}\varepsilon_i(t)}{e_t(X_i)} + \mu_t(X_i)\right|^2 \\ &= \frac{D_{t,i}\varepsilon_i(t)^2}{e_t(X_i)^2} + 2\mu_t(X_i)\frac{D_{t,i}\varepsilon_i(t)}{e_t(X_i)} + \mu_t(X_i)^2 \\ &\leq \sup_{x\in\mathcal{X}} \left|\frac{\pi_t}{e_t(x)}\right|^2 \pi_t^{-2}\varepsilon_i(t)^2 D_{t,i} + 2\sup_{x\in\mathcal{X}} \mu_t(x) \sup_{x\in\mathcal{X}} \left|\frac{\pi_t}{e_t(x)}\right| \pi_t^{-1}|\varepsilon_i(t)|D_{t,i} + \sup_{x\in\mathcal{X}} \mu_t(x)^2 \\ &\lesssim \pi_t^{-2}\varepsilon_i(t)^2 D_{t,i} + \pi_t^{-1}|\varepsilon_i(t)|D_{t,i} + 1\end{aligned}$$

Looking at the max then yields

$$\begin{aligned}
&\sup_{t\neq 0}\max_{1\leq i\leq n}|\psi_i^{[t]}(\eta)|^2 \\
&\quad\lesssim_P \sup_{t\neq 0}\pi_t^{-2}\max_{1\leq i\leq n}|\varepsilon_i(t)|^2\max_{1\leq i\leq n}D_{t,i}+\sup_{t\neq 0}\pi_t^{-1}\max_{1\leq i\leq n}|\varepsilon_i(t)|\max_{1\leq i\leq n}D_{t,i}+1 \\
&\quad\lesssim_P J^2n^{2/m}+Jn^{1/m}+1 \\
&\quad\lesssim J^2n^{2/m}
\end{aligned}$$

Equivalently we obtain

$$\begin{aligned}
\sup_{t\neq 0}\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})| &\leq \sup_{\hat{\eta}\in\hat{\mathcal{H}}_n}\sup_{t\neq 0}\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})| \\
&\lesssim_P \sup_{t\neq 0}\chi_{t,n}\pi_t^{-1}\max_{1\leq i\leq n}|\varepsilon_i(t)|\max_{1\leq i\leq n}D_{t,i} \\
&\lesssim_P Jn^{1/m}
\end{aligned}$$

and

$$\begin{aligned}
\sup_{t\neq 0}\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})|^2 &\leq \sup_{\hat{\eta}\in\hat{\mathcal{H}}_n}\sup_{t\neq 0}\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})|^2 \\
&\lesssim_P \sup_{t\neq 0}\chi_{t,n}\pi_t^{-2}\max_{1\leq i\leq n}|\varepsilon_i(t)|^2\max_{1\leq i\leq n}D_{t,i} \\
&\lesssim_P J^2n^{2/m}
\end{aligned}$$

(MA.2) $\kappa_{n,1}$ and $\kappa_{n,2}$ rates

$$\begin{aligned}
E[\max_{1\leq i\leq n}|\psi_i(\hat{\eta},\pi)-\psi_i(\eta,\pi)|] &\lesssim_P \sum_{t\neq 0}\pi_t E[\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})-\psi_i^{[t]}(\eta)|] \\
&\lesssim_P J\sup_{t\neq 0}\pi_t E[\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})-\psi_i^{[t]}(\eta)|] \\
&\leq \sup_{t\neq 0}E[\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})-\psi_i^{[t]}(\eta)|] \\
&\leq \kappa_{n,1}
\end{aligned}$$

$$
\begin{aligned}
E[\max_{1\le i\le n}|\psi_i(\hat\eta,\pi)-\psi_i(\eta,\pi)|^2] &\lesssim_P \sum_{t\neq 0}\sum_{t'\neq 0}\pi_t\pi_{t'}E[\max_{1\le i\le n}|\psi_i^{[t]}(\hat\eta)-\psi_i^{[t]}(\eta)||\psi_i^{[t']}(\hat\eta)-\psi_i^{[t']}(\eta)|] \\
&\lesssim_P J\sup_{t\neq 0}\pi_t E[\max_{1\le i\le n}|\psi_i^{[t]}(\hat\eta)-\psi_i^{[t]}(\eta)|^2] \\
&\le \sup_{t\neq 0} E[\max_{1\le i\le n}|\psi_i^{[t]}(\hat\eta)-\psi_i^{[t]}(\eta)|^2] \\
&\le \kappa_{n,2}
\end{aligned}
$$

(MA.3) v_i **decomposition** For arbitrary η and π define

$$
\hat\beta(\eta,\pi)=\hat Q^{-1}E_n[b_i\psi_i(\eta,\pi)]
$$

Thus $\hat\beta=\hat\beta(\hat\eta,\hat\pi)$. Rewriting the residual using estimated nuisances then yields

$$
\begin{aligned}
v_i &= \psi_i(\hat\eta,\hat\pi)-b_i'\hat\beta(\hat\eta,\hat\pi) \\
&= \psi_i(\hat\eta,\pi)-b_i'\hat\beta(\hat\eta,\pi)+\psi_i(\hat\eta,\hat\pi)-\psi_i(\hat\eta,\pi)+b_i'(\hat\beta(\hat\eta,\pi)-\hat\beta(\hat\eta,\hat\pi))
\end{aligned}
$$

(MA.4) Unconditional probability rates For any $t\neq 0$ the estimation error of the probability weights is given by

$$
\begin{aligned}
\frac{\hat\pi_t}{1-\hat\pi_0}-\frac{\pi_t}{1-\pi_0} &= \frac{\hat\pi_t(1-\pi_0)-\pi_t(1-\hat\pi_0)}{(1-\hat\pi_0)(1-\pi_0)} \\
&\lesssim_P \Big(\pi_t|\hat\pi_0-\pi_0|+\pi_0|\hat\pi_t-\pi_t|\Big)(1+O_p(|\hat\pi_0-\pi_0|)) \\
&\lesssim_P \pi_t n^{-1/2}+(n/\pi_t)^{-1/2} \\
&\lesssim (Jn)^{-1/2}
\end{aligned}
$$

by Chebyshev's inequality as

$$
\begin{aligned}
E[\hat\pi_t-\pi_t] &= 0 \\
E[(\hat\pi_t-\pi_t)^2] &= \frac{V[D_{it}]}{n}=\pi_t(1-\pi_t)/n\lesssim \pi_t/n
\end{aligned}
$$

and $|\hat\pi_0-\pi_0|\lesssim_P n^{-1/2}$ as control propensities are bounded away from 0 and 1.

(MA.5) Maximal impact of nuisances on predictions

$$\begin{aligned}\max_{1\leq j\leq n}|b_j'(\hat{\beta}(\hat{\eta},\hat{\pi})-\hat{\beta}(\hat{\eta},\pi))| &= \max_{1\leq j\leq n}|b_j'\hat{Q}^{-1}E_n[b_i(\psi_i(\hat{\eta},\hat{\pi})-\psi_i(\hat{\eta},\pi))]| \\ &\lesssim_P \sup_{z\in\mathcal{Z}}||b(z)||\ ||\hat{Q}^{-1}||\ ||E_n[b_i(\psi_i(\hat{\eta},\hat{\pi})-\psi_i(\hat{\eta},\pi))]|| \\ &\lesssim_P \xi_k\sqrt{\frac{Jk}{n}}\end{aligned}$$

where the second to last line comes from H.1 - H.3. Moreover,

$$\begin{aligned}\max_{1\leq j\leq n}|b_j'(\hat{\beta}(\hat{\eta},\pi)-\hat{\beta}(\eta,\pi))| &= \max_{1\leq j\leq n}|b_j'\hat{Q}^{-1}E_n[b_i(\psi_i(\hat{\eta},\pi)-\psi_i(\eta,\pi))]| \\ &\lesssim_P \xi_k||\hat{Q}^{-1}||\ ||E_n[b_i(\psi_i(\hat{\eta},\pi)-\psi_i(\eta,\pi))]|| \\ &\lesssim_P \xi_k n^{-1/2}(B_n^{[rATE]}+\Lambda_n^{[rATE]}) \\ &\lesssim_P \xi_k n^{-1/2}\end{aligned}$$

due to Markov's inequality. The deviation from the best linear predictor follows from BCCK, Theorem 4.3, under the assumptions stated in the beginning of this section:

$$\max_{1\leq j\leq n}|b_j'(\hat{\beta}(\eta,\pi)-\beta_0)| \lesssim_P \xi_k\sqrt{\frac{\log(k)}{n}}$$

(MA.6) Error term tail bounds The marginal tail bound C.1 imply the following tail bound for $rATE$

$$\begin{aligned}\max_{1\leq i\leq n}|\varepsilon_i| &= \max_{1\leq i\leq n}|\psi_i^{[rATE]}(\eta,\pi)-E[\psi_i^{[rATE]}(\eta,\pi)|Z_i]| \\ &\lesssim_P \max_{1\leq i\leq n}\sum_{t\neq 0}\frac{\pi_t D_{t,i}|\varepsilon_i(t)|}{e_t(X_i)} \\ &\lesssim_P \max_{1\leq i\leq n}\sup_{t\neq 0}\sup_{x\in\mathcal{X}}\frac{\pi_t}{e_t(x)}|\varepsilon_i(t)|\sum_{t\neq 0}D_{t,i} \\ &\lesssim_P \max_{1\leq i\leq n}\sup_{t\neq 0}|\varepsilon_i(t)|\end{aligned}$$

and equivalently by B.1 for the $nATE$ with $\sup_{x\in\mathcal{X}} \frac{\pi_t}{e_t(x)}$ replaced by 1. Note that

$$\begin{aligned}
E[\max_{1\leq i\leq n}\sup_{t\neq 0}|\varepsilon_i(t)|] &= \int_{-\infty}^{(nJ)^{1/m}} P(\max_{1\leq i\leq n}\sup_{t\neq 0}|\varepsilon_i(t)|>w)dw \\
&\quad + \int_{(nJ)^{1/m}}^{\infty} P(\max_{1\leq i\leq n}\sup_{t\neq 0}|\varepsilon_i(t)|>w)dw \\
&\leq (Jn)^{1/m} + \int_{(nJ)^{1/m}}^{\infty} n\sum_{t\neq 0} P(|\varepsilon_i(t)|>w)dw \\
&\leq (Jn)^{1/m} + (Jn)^{1/m}\int_{(nJ)^{1/m}}^{\infty} w^{1-1/m}P(|\varepsilon_i(t)|>w)dw \\
&\lesssim (Jn)^{1/m}(1+o(1)) \\
&\lesssim (Jn)^{1/m}
\end{aligned}$$

where the second term is convergent due to the conditional moment bound B.1/C.1 for $\varepsilon_i(t)$. Overall, this implies that $\max_{1\leq i\leq n}|\varepsilon_i| \lesssim_P (Jn)^{1/m}$ by Markov's inequality.

B.6.2 Definitions and Decomposition

Define

$$\begin{aligned}
\Sigma &= E[(b_i(\varepsilon_i + r_i) - \gamma a_i)(b_i(\varepsilon_i + r_i) - \gamma a_i)'] \\
\Sigma_n &= E_n[(b_i(\varepsilon_i + r_i) - \gamma a_i)(b_i(\varepsilon_i + r_i) - \gamma a_i)'] \\
\hat{\Sigma}_n &= E_n[(b_i v_i - \hat{\gamma}\hat{a}_i)(b_i v_i - \hat{\gamma}\hat{a}_i)']
\end{aligned}$$

with $v_i = \psi_i(\hat{\eta}, \hat{\pi}) - b_i'\hat{\beta}$ and $\hat{a}_i$ obtained by replacing the true probabilities π in a_i with the sample estimates $\hat{\pi}$:

$$\begin{aligned}
\hat{a}_i &= (\hat{a}_i^{[1]}, \ldots, \hat{a}_i^{[J]}) \\
\hat{a}_i^{[t]} &= \frac{D_{t,i}(1-\hat{\pi}_0) + D_{0,i}\hat{\pi}_t - \hat{\pi}_t}{(1-\hat{\pi}_0)^2}
\end{aligned}$$

In the following we proof that $||\hat{\Sigma}_n - \Sigma|| = o_p(1)$. The remaining rates for convergence of $\hat{\Omega} = \hat{Q}^{-1}\hat{\Sigma}_n\hat{Q}^{-1}$ to Ω can then be obtained directly from BCCK, Proof of Theorem 4.6.

First note the general decomposition

$$||\hat{\Sigma}_n - \Sigma|| \leq ||\hat{\Sigma}_n - \Sigma_n|| + ||\Sigma_n - \Sigma||$$

B.6.3 $||\hat{\Sigma}_n - \Sigma_n||$

Consider the decomposition:

$$\begin{aligned} ||\hat{\Sigma}_n - \Sigma_n|| \leq & ||E_n[b_i b_i'(v_i^2 - (\varepsilon_i + r_i)^2)]|| \\ & + ||E_n[\hat{\gamma}\hat{a}_i\hat{a}_i'\hat{\gamma}' - \gamma a_i a_i'\gamma']|| \\ & + ||E_n[b_i(v_i(\hat{\gamma}\hat{a}_i)' - (\varepsilon_i + r_i)(\gamma a_i))]|| \end{aligned}$$

We bound the three components separately.

Part 1

$$\begin{aligned} & ||E_n[b_i b_i'(v_i^2 - (\varepsilon_i + r_i)^2)]|| \\ & \leq ||E_n[b_i b_i'((\psi_i(\hat{\eta}, \pi) - b_i'\hat{\beta}(\hat{\eta}, \pi) + \psi_i(\hat{\eta}, \hat{\pi}) - \psi_i(\hat{\eta}, \pi) + b_i'(\hat{\beta}(\hat{\eta}, \pi) - \hat{\beta}(\hat{\eta}, \hat{\pi})))^2 - (\varepsilon_i + r_i)^2)]|| \\ & \leq ||E_n[b_i b_i'(\psi_i(\hat{\eta}, \pi) - b_i'\hat{\beta}(\hat{\eta}, \pi) - (\varepsilon_i + r_i)^2)]|| + ||E_n[b_i b_i'(\psi_i(\hat{\eta}, \hat{\pi}) - \psi_i(\hat{\eta}, \pi))^2]|| \\ & \quad + ||E_n[b_i b_i'(b_i'(\hat{\beta}(\hat{\eta}, \pi) - \hat{\beta}(\hat{\eta}, \hat{\pi})))^2]|| \\ & \equiv (v.1) + (v.2) + (v.3) \end{aligned}$$

For $(v.1)$ we can use the same proof as in SC, Theorem 3.3 for the $nATE$ and obtain.

$$(v.1) \lesssim_P \left((Jn)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)||\right)\kappa_{n,1} + \kappa_{n,2}$$

due to (MA.6). Note that for $nATE$, $rATE$, and Δ, the tail bounds and approximation errors are allowed to differ. For the second term we have

$$\begin{aligned}
(v.2) &= \left|\left| E_n\left[b_i b_i' \left(\sum_{t\neq 0} \psi_i^{[t]}(\hat{\eta}) \left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0} \right) \right)^2 \right] \right|\right| \\
&= \left|\left| \sum_{t\neq 0} \sum_{t'\neq 0} E_n\left[b_i b_i' \psi_i^{[t]}(\hat{\eta}) \psi_i^{[t']}(\hat{\eta}) \left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0} \right) \left(\frac{\hat{\pi}_{t'}}{1-\hat{\pi}_0} - \frac{\pi_{t'}}{1-\pi_0} \right) \right) \right] \right|\right| \\
&\lesssim_P J^2 \sup_{t,t'\neq 0} |\psi_i^{[t]}(\hat{\eta}) \psi_i^{[t']}(\hat{\eta})| |E_n[b_i b_i']| | \frac{\sqrt{\pi_t \pi_t'}}{n} \\
&\lesssim_P J^2 \sup_{t\neq 0} \pi_t^{-2} \max_{1\leq i\leq n} |\varepsilon_i(t)|^2 ||\hat{Q}|| \frac{\pi_t}{n} \\
&\lesssim_P J^3 n^{-(1-2/m)}
\end{aligned}$$

For the third term we have

$$\begin{aligned}
(v.3) &= ||E_n[b_i b_i (b_i' \hat{Q}^{-1} E_n[b_i'(\psi(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))])^2]|| \\
&\lesssim_P \max_{1\leq j\leq n} |b_j' \hat{Q}^{-1} E_n[b_i'(\psi(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))]|^2 ||\hat{Q}|| \\
&\lesssim_P \sup_{z\in\mathcal{Z}} ||b(z)||^2 ||\hat{Q}^{-1}||^2 ||E_n[b_i'(\psi(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))]||^2 ||\hat{Q}|| \\
&\lesssim_P \xi_k^2 \left(\sqrt{\frac{kJ}{n}} \right)^2 \\
&= \frac{\xi_k^2 kJ}{n}
\end{aligned}$$

where the second to last line comes from (H.3).

Part 2 Now we use (H.5) to bound the second variance term. Let $A_n = E_n[a_i a_i']$ and $\hat{A}_n = E_n[\hat{a}_i \hat{a}_i']$. We have that

$$
\begin{aligned}
&||E_n[\hat{\gamma}\hat{a}_i\hat{a}_i'\hat{\gamma}' - \gamma a_i a_i'\gamma']|| \\
&\quad = ||\hat{\gamma}\hat{A}_n\hat{\gamma}' - \gamma A_n \gamma'|| \\
&\quad \leq ||\hat{\gamma}-\gamma||^2||\hat{A}_n - A_n|| + ||\hat{\gamma}-\gamma||^2||A_n|| + ||\hat{\gamma}-\gamma||\ ||\hat{A}_n - A_n||\ ||\gamma|| \\
&\qquad + ||\hat{\gamma}-\gamma||\ ||A_n||\ ||\gamma|| + ||\hat{A}_n - A_n||\ ||\gamma||^2 \\
&\quad \lesssim_P ||\hat{\gamma}-\gamma||\ ||A_n||\ ||\gamma|| + ||\hat{A}_n - A_n||\ ||\gamma||^2 \\
&\quad \lesssim_P \sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2})\sqrt{kJ} + n^{-1/2}kJ \\
&\quad \lesssim \sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2})
\end{aligned}
$$

The covariance term is bounded by

$$
\begin{aligned}
&||E_n[b_i(v_i(\hat{\gamma}\hat{a}_i)' - (\varepsilon_i + r_i)(\gamma a_i))]|| \\
&\quad \leq ||E_n[b_i(v_i - (\varepsilon_i + r_i))(\gamma a_i)']|| + ||E_n[b_i(\varepsilon_i + r_i)(\hat{\gamma}\hat{a}_i - \gamma a_i)']|| \\
&\qquad + ||E_n[b_i(v_i - (\varepsilon_i + r_i))(\hat{\gamma}\hat{a}_i - \gamma a_i)']|| \\
&\quad \equiv (c.1) + (c.2) + (c.3)
\end{aligned}
$$

Now we use (H.5) and (MA.5) to bound $(c.1)$

$$
\begin{aligned}
&||E_n[b_i(v_i - (\varepsilon + r_i))(\gamma a_i)']|| \\
&\quad \lesssim_P \xi_k \max_{1\leq i\leq n} |b_i'(\hat{\beta}(\hat{\eta},\hat{\pi}) - \beta_0)| E_n[||\gamma a_i||] \\
&\qquad + ||E_n[b_i(\psi_i(\hat{\eta},\hat{\pi}) - \psi_i(\hat{\eta},\pi))(\gamma a_i)']|| + ||E_n[b_i(\psi_i(\hat{\eta},\pi) - \psi_i(\eta,\pi))(\gamma a_i)']|| \\
&\quad \equiv (c.1.1) + (c.1.2) + (c.1.3)
\end{aligned}
$$

with decomposing the BLP error using the rates in (H.5) and (MA.5)

$$\begin{aligned}(c.1.1) &\leq \xi_k J||\gamma||E_n[||a_i||] \max_{1\leq i\leq n} |b_i'(\hat{\beta}(\hat{\eta},\hat{\pi}) - \beta_0)| \\ &\lesssim_P \xi_k\sqrt{kJ}\left(\xi_k\sqrt{Jk/n} + \xi_k n^{-1/2} + \xi_k\sqrt{\log(k)/n}\right) \\ &\lesssim \frac{\xi_k^2 kJ}{\sqrt{n}}\end{aligned}$$

For $(c.1.2)$ note that

$$\begin{aligned}(c.1.2) &= ||\sum_{t\neq 0}\left(\frac{\hat{\pi}_t}{1-\hat{\pi}_0} - \frac{\pi_t}{1-\pi_0}\right)E_n[b_i\psi_i^{[t]}(\hat{\eta})a_i^{[t]}]\gamma_t'|| \\ &\lesssim_P J\sup_{t\neq 0}\sqrt{\frac{\pi_t}{n}}\max_{1\leq i\leq n}|\psi_i^{[t]}(\hat{\eta})|E_n[||b_i a_i^{[t]}||]||\gamma_t|| \\ &\lesssim_P J^{3/2}n^{-(\frac{1}{2}-\frac{1}{m})}k\end{aligned}$$

by (MA.1) and (H.5) in conjunction with Markov's inequality. For $(c.1.3)$, we have that

$$\begin{aligned}(c.1.3) &= ||\sum_{t\neq 0}\frac{\pi_t}{1-\pi_0}E_n[b_i(\psi_i^{[t]}(\hat{\eta}) - \psi_i^{[t]}(\eta))a_i^{[t]}]\gamma_t'|| \\ &\leq J\sup_{t\neq 0}\pi_t\max_{1\leq j\leq n}|\psi_j^{[t]}(\hat{\eta}) - \psi_j^{[t]}(\eta)||a_i^{[t]}|E_n[||b_i||]||\gamma_t|| \\ &\lesssim_P \kappa_{1,n}k\end{aligned}$$

also by (H.5) and Markov's inequality. For $(c.2)$ note that

$$\begin{aligned}(c.2) &\leq \xi_k\max_{1\leq j\leq n}|\varepsilon_j + r_j|E_n[||\hat{\gamma}\hat{a}_i - \gamma a_i||] \\ &\lesssim_P \xi_k((Jn)^{1/m} + \sup_{z\in\mathcal{Z}}||r(z)||)E_n[||\hat{\gamma}\hat{a}_i - \gamma a_i||] \\ &\lesssim_P \xi_k((Jn)^{1/m} + \sup_{z\in\mathcal{Z}}||r(z)||)\sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2})\end{aligned}$$

as

$$\begin{aligned}
E_n[||\hat{\gamma}\hat{a}_i - \gamma a_i||] &= E_n[||\hat{\gamma} - \gamma||\ ||a_i||] + E_n[||\gamma||\ ||\hat{a}_i - a_i||] + E_n[||\hat{\gamma} - \gamma||\ ||\hat{a}_i - a_i||] \\
&\lesssim_P ||\hat{\gamma} - \gamma|| E_n[||a_i||] + ||\gamma|| E_n[||\hat{a}_i - a_i||] \\
&\lesssim_P \sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2}) + \sqrt{Jk/n} \\
&\lesssim \sqrt{kJ^2}(n^{-1/2} + m_{n,2} + Js_{n,2})
\end{aligned}$$

by (H.5). Now note that $(c.3)$ is at most of rate $(c.1)+(c.2)$ which completes the covariance part.

B.6.4 $||\Sigma_n - \Sigma||$

Now consider the remaining difference

$$\begin{aligned}
||\Sigma_n - \Sigma|| &\leq ||E_n[b_i b_i'(\varepsilon_i + r_i)] - E[b_i b_i'(\varepsilon_i + r_i)]|| + ||E_n[\gamma a_i a_i'\gamma'] - E[\gamma a_i a_i'\gamma']|| \\
&\quad + ||E_n[b_i(\varepsilon_i + r_i)(\gamma a_i)' + \gamma a_i(\varepsilon_i + r_i)b_i'] - E[b_i(\varepsilon_i + r_i)(\gamma a_i)' + \gamma a_i(\varepsilon_i + r_i)b_i']|| \\
&= (d.1) + (d.2) + (d.3)
\end{aligned}$$

$(d.1)$ is bounded in BCCK, Proof of Theorem 4.6, with adapted tail rates using (MA.6)

$$(d.1) \lesssim_P \left((Jn)^{1/m} + \sup_{z\in\mathcal{Z}} ||r(z)||\right)\sqrt{\frac{\xi_k^2 \log k}{n}}$$

For $(d.2)$ note that

$$\begin{aligned}
(d.2) &\leq ||\gamma||^2 ||E_n[a_i a_i'] - E[a_i a_i']|| \\
&\lesssim_P kJ\sqrt{\frac{\log J}{n}}
\end{aligned}$$

due to (H.5). For $(d.3)$ first note that

$$\begin{aligned}
\max_{1\leq i\leq n} |a_i'\gamma' b_i| &\lesssim_P \xi_k ||\gamma|| \max_{1\leq i\leq n} ||a_i|| \lesssim_P \xi_k\sqrt{k}J \\
E_n[||b_i a_i'\gamma'||] &\leq \xi_k ||\gamma|| E_n[||a_i||] \lesssim_P \xi_k\sqrt{kJ}
\end{aligned}$$

due to (H.5). Now we use the Symmetrization Lemma to bound $(d.3)$. Let w_i for $i = 1, \ldots, n$ denote independent Rademacher random variables independent of the data. Denote $E_w[\cdot]$ the expectation operator with respect to the measure of w. We have that

$$
\begin{aligned}
(d.3) &\lesssim E\left[E_w\left[||E_n[w_i(\varepsilon_i + r_i)(b_i a_i \gamma' + \gamma a_i b_i')]||\right]\right] \\
&\lesssim \sqrt{\frac{\log k}{n}} E\left[||E_n[(\varepsilon_i + r_i)^2 (b_i a_i' \gamma' b_i a_i' \gamma' + b_i a_i' \gamma' \gamma a_i b_i' + \gamma a_i b_i' b_i a_i' \gamma' + \gamma a_i b_i' \gamma a_i b_i')]||^{1/2}\right] \\
&\lesssim_P \sqrt{\frac{\log k}{n}} \max_{1 \le i \le n} |\varepsilon_i + r_i| \Bigg(2 \max_{1 \le i \le n} |a_i \gamma b_i|^{1/2} E[E_n[||b_i a_i' \gamma'||]]^{1/2} \\
&\quad + \max_{1 \le i \le n} |a_i' \gamma' \gamma a_i|^{1/2} E[E_n[||b_i b_i'||]]^{1/2} + \max_{1 \le i \le n} |b_i' b_i|^{1/2} E[E_n[||\gamma a_i a_i' \gamma'||]]^{1/2} \Bigg) \\
&\lesssim_P \sqrt{\frac{\log k}{n}} \max_{1 \le i \le n} |\varepsilon_i + r_i| \left((\xi_k \sqrt{k} J)^{1/2} (\xi_k \sqrt{kJ})^{1/2} + ||\gamma|| \max_{1 \le i \le n} ||a_i|| \sqrt{k} + \xi_k ||\gamma|| \right) \\
&\lesssim \sqrt{\frac{\log k}{n}} \left((nJ)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)|| \right) \left(J^{3/4} \xi_k \sqrt{k} + kJ \right)
\end{aligned}
$$

where the second line is due to Khinchin's inequality as $(b_i a_i \gamma' + \gamma a_i b_i')$ are iid symmetric. The remaining steps follow from (H.5) and (MA.6) and repeated application of Markov's inequality.

B.6.5 Full variance

Collecting all the rates, we obtain that

$$
\begin{aligned}
||\hat{\Sigma}_n - \Sigma|| \lesssim_P & \left((Jn)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)|| \right) \kappa_{n,1} + \kappa_{n,2} + J^3 n^{-(1-2/m)} + \frac{\xi_k^2 k J}{n} \\
& + \frac{\xi_k^2 k J}{\sqrt{n}} + J^{3/2} n^{-(\frac{1}{2} - \frac{1}{m})} k + \kappa_{1,n} k \\
& + \xi_k ((Jn)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)||) \sqrt{kJ^2} (n^{-1/2} + m_{n,2} + J s_{n,2}) \\
& + \left((Jn)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)|| \right) \sqrt{\frac{\xi_k^2 \log k}{n}} + kJ \sqrt{\frac{\log J}{n}} \\
& + \sqrt{\frac{\log k}{n}} \left((nJ)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)|| \right) \left(J^{3/4} \xi_k \sqrt{k} + kJ \right) \\
& \equiv M_{n,2} \qquad (19)
\end{aligned}
$$

For the $nATE$ the derivations are analogue, but there are no γa_i terms. Thus the solution simplifies to

$$||\hat{\Sigma}_n - \Sigma|| \lesssim_P \left((Jn)^{1/m} + \sup_{z \in \mathcal{Z}} ||r(z)||\right)\left(\kappa_{n,1} + \sqrt{\frac{\xi_k^2 \log k}{n}}\right) + \kappa_{n,2} = M_{n,1} \tag{20}$$

Thus assuming $M_{n,1} = o(1)$ for the $nATE$ or $M_{n,2} = o(1)$ for $rATE/\Delta$ respectively corresponds to Assumption A.V.

B.7 Supplementary Material for Section 6

We simulate n observations of (Y_i, X_i, T_i). Let X_i be a k-dimensional vector of uniform random variables $X_{i,j} \sim \mathcal{U}[-1, 1]$ for $j = 1, \ldots, p$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. We let $Y_i(t) = u_i$ for $t \neq 1$ and $Y_i(1) = \tau + u_i$. Treatment probabilities $P(T_i = t|X_i) = e_t(X_i)$ for $t = 0, 1 \ldots, J$ (with $t = 0$ denoting control) are generated under independence of irrelevant alternatives as

$$e_0(x) = \frac{1}{1 + \sum_{j \neq 0} \exp(x_1 \beta_j)}$$
$$e_t(x) = \frac{\exp(x_1 \beta_t)}{1 + \sum_{j \neq 0} \exp(x_1 \beta_j)}$$

with $\beta_1 = 1$ and $\beta_t = 0$ for all $t \neq 1$. Thus, conditional treatment effects are given by $\tau_1(x) = \tau = 10$ and $\tau_t(x) = 0$ for all $t \neq 1$. This implies the following conditional decomposition terms (II):

$$E[rATE(X_i)|X_{i,1} = x_1] = \tau\left[\frac{\pi_1}{1 - \pi_0}\right]$$
$$E[nATE(X_i)|X_{i,1} = x_1] = \tau\left[\frac{e_1(x_1)}{1 - e_0(x_1)}\right]$$
$$E[\Delta(X_i)|X_{i,1} = x_1] = \tau\left[\frac{e_1(x_1)}{1 - e_0(x_1)} - \frac{\pi_1}{1 - \pi_0}\right]$$

Note that $E[X_{i,1}] = 0$ and $V[X_{i,1}] = 1/3$. Thus the best linear approximation of $E[\Delta(X_i)|X_{i,1}]$ has population parameters (α, β) with

$$\begin{aligned}
\alpha &= \tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} - \frac{\pi_1}{1-\pi_0}\right] - \beta E[X_{i,1}] \\
&= \tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} - \frac{\pi_1}{1-\pi_0}\right] \\
\beta &= \frac{\tau}{V[X_{i,1}]} E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} X_{i,1} - \frac{\pi_1}{1-\pi_0} X_{i,1}\right] \\
&= 3\tau E\left[\frac{e_1(X_{i,1})}{1-e_0(X_{i,1})} X_{i,1}\right]
\end{aligned}$$

and equivalently for the $rATE$ and $nATE$. Evaluating the expectation yields the following parameterization:

Table B.1: Monte Carlo Study: Parameterization

	$rATE$	$nATE$	Δ
α	5.127	5.000	-.127
β	0.000	2.383	2.383

B.8 Supplementary Material for Section 7.1

The distribution of smoking intensities is shown in Figure B.4 and Table B.2. The majority of mothers do not smoke during pregnancy ranging from 76% for Black mothers to 96% in the category "Other". However, the right panel of Figure B.4 shows that conditional on smoking white mothers and older mothers smoke more heavily.

Table B.2: Distribution of smoking intensities by ethnicity (in percent)

	Black	Hispanic	Other	White	All
$>$ 20 cigs	0.7	0.5	0.2	1.2	1.1
16-20 cigs	4.4	2.7	0.9	5.5	5.1
11-15 cigs	0.7	0.5	0.2	1.3	1.2
6-10 cigs	11.5	5.4	1.5	7.4	7.8
1-5 cigs	6.7	4.0	1.2	3.0	3.6
None	76.1	87.0	96.0	81.6	81.2

Figure B.5 replicates the solid line of Figure 1 in Cattaneo (2010) with Double Machine Learning as a byproduct. Our results are very similar and show that average potential outcomes become smaller the higher the intensity of smoking.

Figure B.4: Distribution of smoking intensities along heterogeneity variables

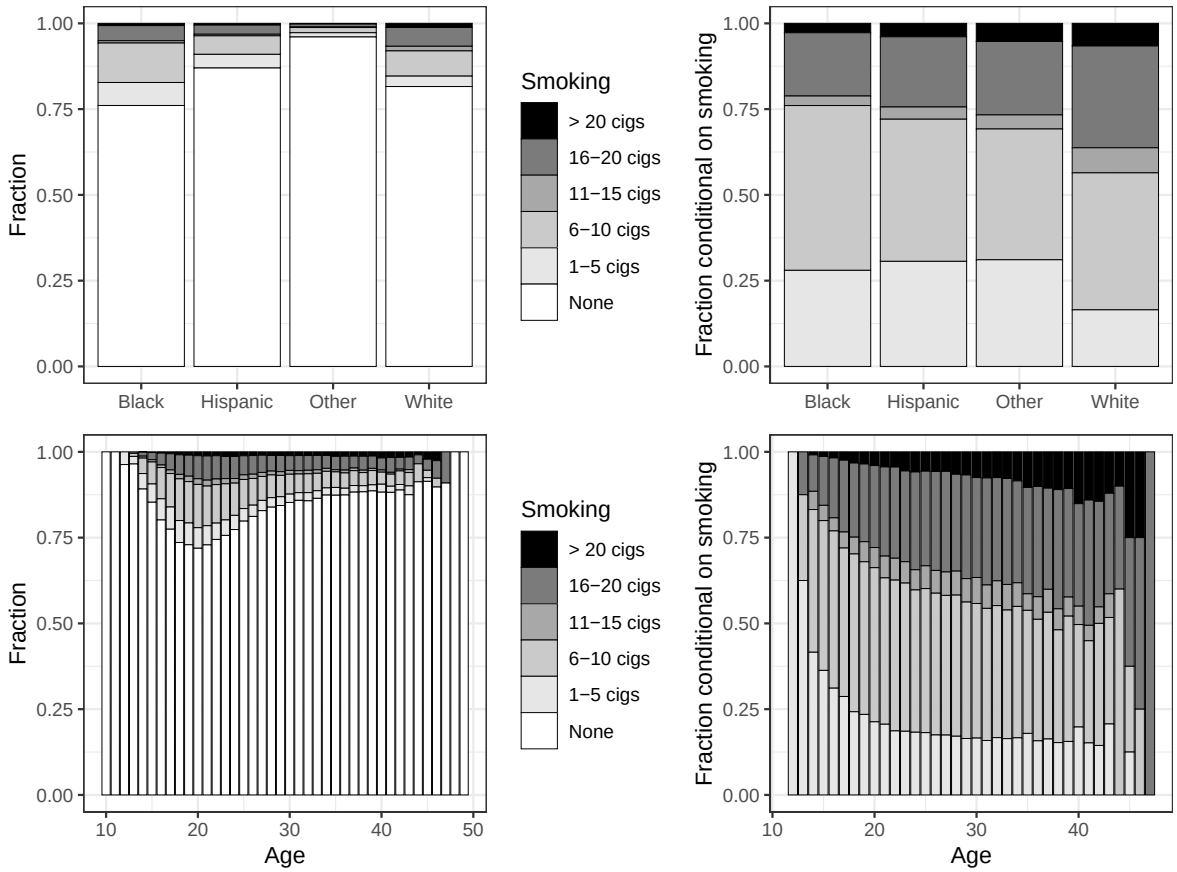


Figure B.6 contains the re-scaled propensity scores $e_t(x)/\pi_t$ for all treatment versions.

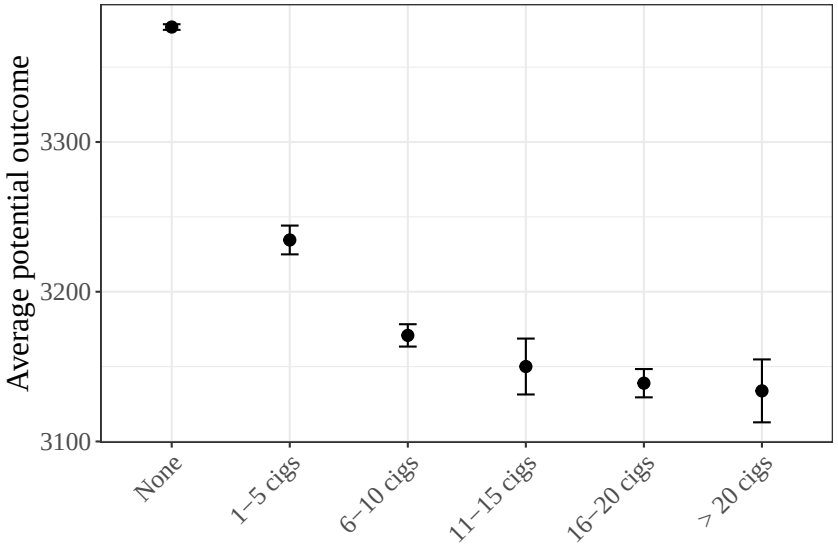
B.9 Supplementary Material for Section 7.2

The distribution of versions is shown in Figure B.7 and Table B.3. We observe that women are overrepresented in clerical, health and food training, while men are more likely to be observed in automechanics, welding, electrical and construction training.

As a byproduct of the decomposition estimation, we create the AIPW scores for every treatment version. This allows us to inspect their often noisily estimated average potential outcomes in Figure B.8. We observe a clear pattern. The point estimates of the predominantly male trainings are all larger than the predominantly female ones.

Figure B.9 contains the re-scaled propensity scores $e_t(x)/\pi_t$ for all treatment versions.

Figure B.5: Average potential outcomes of smoking intensities



Note: Average potential outcomes estimated with Double Machine Learning using an ensemble of Ridge, Lasso and Random Forest regression. Point estimates and 95%-confidence interval.

Figure B.6: Re-scaled propensity scores

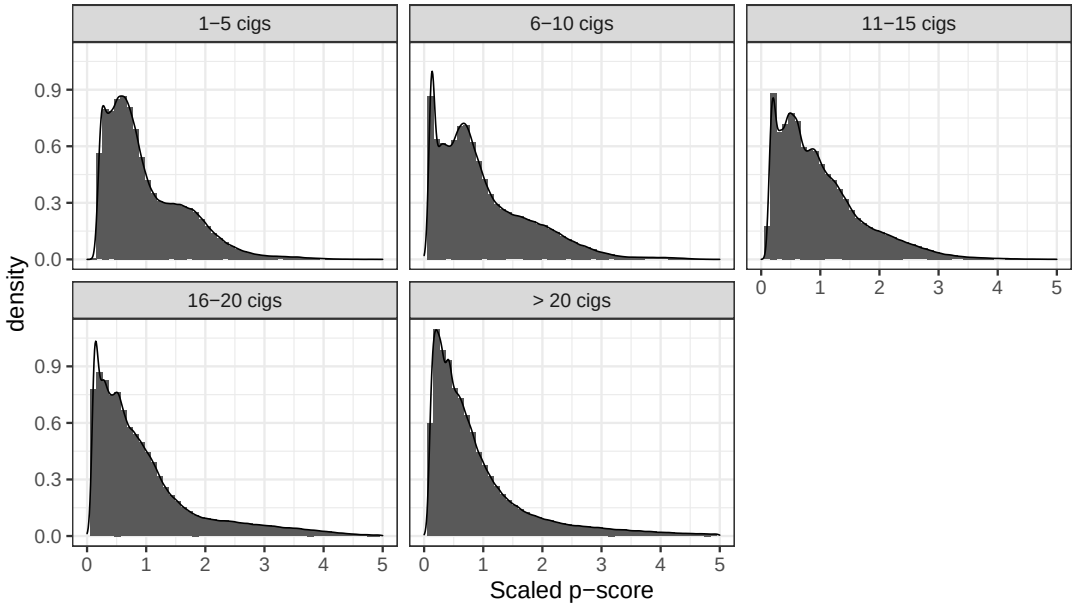


Figure B.7: Distribution of treatment versions by gender

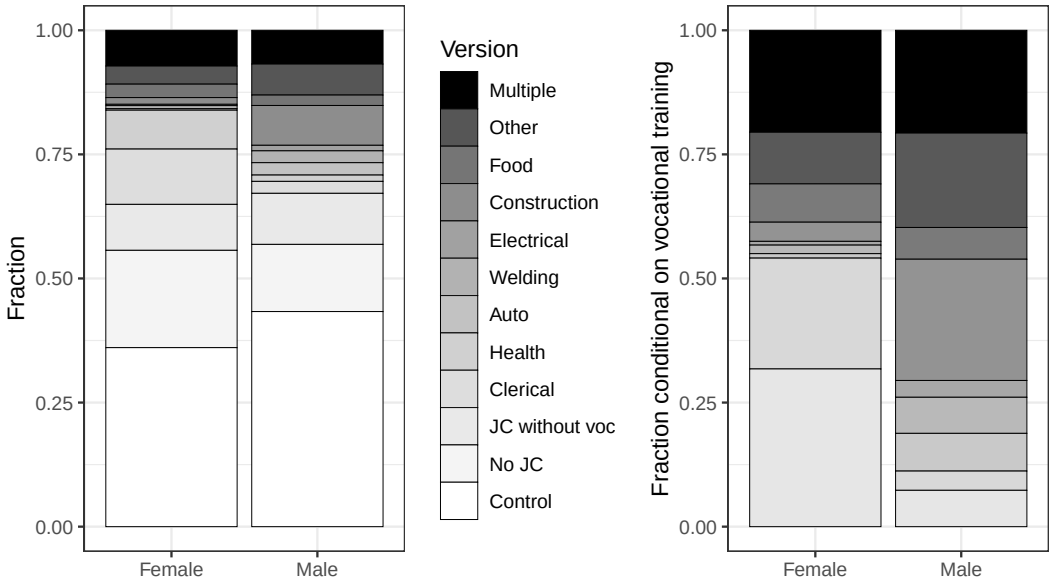
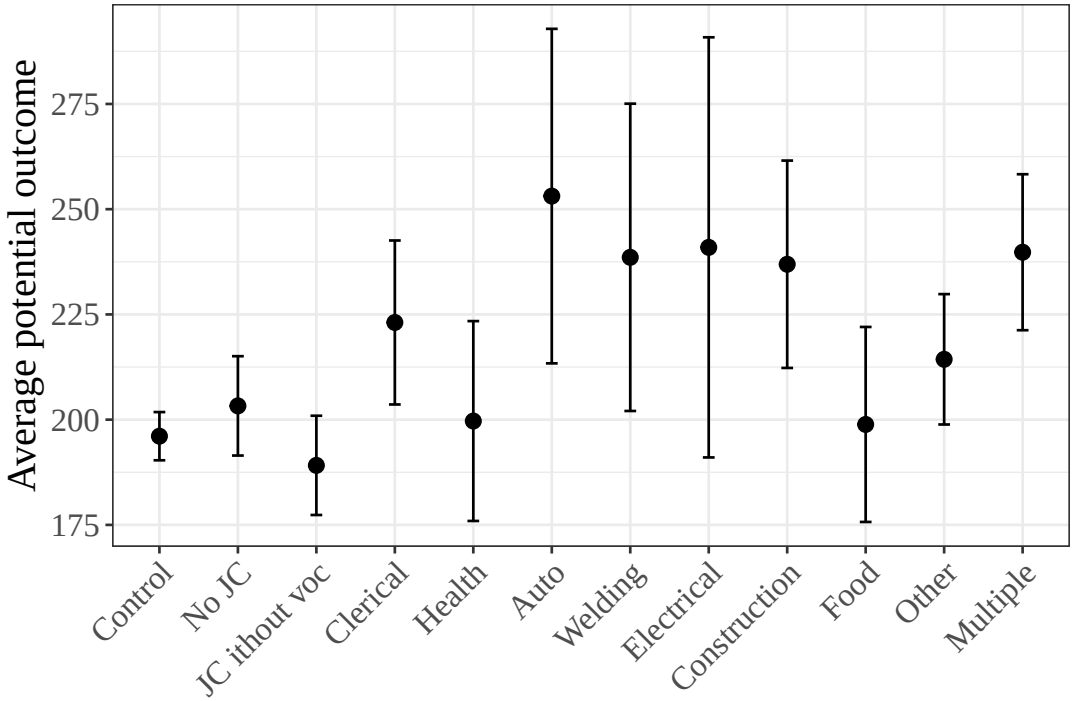


Table B.3: Share of observations in treatment versions (in percent)

	Female	Male	All
Control	36.1	43.3	40.2
No JC	19.6	13.5	16.1
JC without voc	9.3	10.3	9.8
Clerical	11.1	2.4	6.1
Health	7.8	1.3	4.1
Auto	0.3	2.5	1.6
Welding	0.6	2.4	1.6
Electrical	0.3	1.1	0.8
Construction	1.4	8.0	5.2
Food	2.7	2.1	2.4
Other	3.6	6.2	5.1
Multiple	7.2	6.8	7.0

Figure B.8: Average potential outcomes of treatment versions



Note: Average potential outcomes estimated with Double Machine Learning using an ensemble of Ridge, Lasso and Random Forest regression. Point estimates and 95%-confidence interval.

Figure B.9: Re-scaled propensity scores