Double machine learning for sample selection models

Michela Bia*, Martin Huber**, and Lukáš Lafférs+

*Luxembourg Institute of Socio-Economic Research and University of Luxembourg

**University of Fribourg, Dept. of Economics and

Center for Econometrics and Business Analytics, St. Petersburg State University

+Matej Bel University, Dept. of Mathematics

Abstract: This paper considers the evaluation of discretely distributed treatments when outcomes are only observed for a subpopulation due to sample selection or outcome attrition. For identification, we combine a selection-on-observables assumption for treatment assignment with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. We also consider dynamic confounding, meaning that covariates that jointly affect sample selection and the outcome may (at least partly) be influenced by the treatment. To control in a data-driven way for a potentially high dimensional set of pre- and/or post-treatment covariates, we adapt the double machine learning framework for treatment evaluation to sample selection problems. We make use of (a) Neyman-orthogonal, doubly robust, and efficient score functions, which imply the robustness of treatment effect estimation to moderate regularization biases in the machine learning-based estimation of the outcome, treatment, or sample selection models and (b) sample splitting (or cross-fitting) to prevent overfitting bias. We demonstrate that the proposed estimators are asymptotically normal and root-n consistent under specific regularity conditions concerning the machine learners and investigate their finite sample properties in a simulation study. We also apply our proposed methodology to the Job Corps data for evaluating the effect of training on hourly wages which are only observed conditional on employment. The estimator is available in the *causalweight* package for the statistical software R.

Keywords: sample selection, double machine learning, doubly robust estimation, efficient score.

JEL classification: C21.

We have benefited from comments by Alyssa Carlson, David Kaplan, Peter Mueser, and seminar participants at the University of Missouri.

Addresses for correspondence: Michela Bia, Luxembourg Institute of Socio-Economic Research, 11 Porte des Sciences, Maison des Sciences Humaines, 4366 Esch-sur-Alzette/Belval, Luxembourg, michela.bia@liser.lu, michela.bia@ext.uni.lu; Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland, martin.huber@unifr.ch; Lukáš Lafférs, Matej Bel University, Tajovskeho 40, 97411 Banská Bystrica, Slovakia, lukas.laffers@gmail.com. Lafférs acknowledges support provided by the Slovak Research and Development Agency under contract no. APVV-17-0329 and VEGA-1/0692/20.

1 Introduction

In many studies aiming at evaluating the causal effect of a treatment or policy intervention, the empirical analysis is complicated by non-random outcome attrition or sample selection. Examples include the estimation of the returns to education when wages are only observed for the selective subpopulation of working individuals or the effect of educational interventions like vouchers for private schools on college admissions tests when students non-randomly abstain from the test. Furthermore, in observational studies, treatment assignment is typically not random, implying that the researcher faces a double selection problem, namely selection into the treatment and observability of the outcome. A large literature addresses treatment selection by assuming a selection-on-observables assumption, implying that treatment is as good as randomly assigned conditional on observed pre-treatment covariates, see for instance the reviews by Imbens (2004) and Imbens and Wooldridge (2009). Furthermore, a growing number of studies addresses the question of how to control for the crucial confounders in a potentially high-dimensional vector of covariates in a data-driven way based on machine learning algorithms, see for instance the double machine learning framework of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

In this paper, we adapt the double machine learning framework to the evaluation of binary or multiply discrete treatments in the presence of sample selection or outcome attrition. In terms of identifying assumptions, we combine a selection-on-observables assumption for the treatment assignment with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. Such assumptions have previously been considered in Huber (2012) and Huber (2014b) for the estimation of the average treatment effect (ATE) based on inverse probability weighting, however, for pre-selected (or fixed) covariates. As methodological advancement, we derive doubly robust and efficient score functions for evaluating treatment effects under double selection and demonstrate that they satisfy so-called Neyman (1959) orthogonality. The latter property permits controlling for covariates in a data-driven way by machine learning-based estimation of the treatment, outcome, and attrition models under specific conditions. Therefore, the subset of important confounders need not be known a priori (but must be contained in the total set of covariates), which is particularly useful in high dimensional data with a vast number of covariates that could potentially serve as control variables. We

also consider dynamic confounding based on a sequential selection-on-observables assumption that is closely related to assumptions found in the dynamic treatment effect literature as e.g. in Robins (1986), Robins (1998), and Lechner (2009). This assumption permits that covariates that jointly affect sample selection and the outcome may themselves be a function of the treatment, a scenario widely neglected in sample selection models despite its likely relevance in empirical applications. In particular when there is a substantial time lag between treatment assignment and the sample selection process, exploiting post-treatment covariates to tackle selection-outcome confounding seems more convincing than solely relying on pre-treatment covariates (as in conventional selection-on-observables assumptions) for addressing both treatment endogeneity and sample selection.

Following Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), we show that treatment effect estimation based on our score functions (that are tailored to the various identifying assumptions) is root-n consistent and asymptotically normal under particular regularity conditions, in particular the $n^{-1/4}$-convergence of the machine learners. A further condition in the double machine learning framework is the prevention of overfitting bias due to correlations between the various estimation steps. This is obtained by estimating the treatment, outcome, and selection models on the one hand and the treatment effect on the other hand in different parts of the data. As in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), we subsequently swap the roles of the data parts and average over treatment effects in order to prevent asymptotic efficiency losses, a procedure known as cross-fitting. We also provide a simulation study suggesting that our estimators perform decently in terms of the root mean squared error and coverage (by confidence intervals) in the simulation designs with several thousand observations considered. Finally, we present an empirical illustration considering the female sample of a study on Job Corps, a large training program for disadvantaged youth in the U.S. We apply our DML estimators to assess the effects of academic and vocational training on hourly wage, which is only observed conditional on employment, one and four years after program assignment and find some statistical evidence for positive longer-run impacts.

Our paper is related to a range of studies tackling sample selection and selective outcome attrition. One strand of the literature models the attrition process based on a selection-on-observables assumption also known as missing at random (MAR) condition. The latter imposes conditional independence of sample selection and the outcome given observed information like the

covariates and the treatment. Examples include Rubin (1976), Little and Rubin (1987), Carroll, Ruppert, and Stefanski (1995), Shah, Laird, and Schoenfeld (1997), Fitzgerald, Gottschalk, and Moffitt (1998), Abowd, Crepon, and Kramarz (2001), Wooldridge (2002), and Wooldridge (2007). Robins, Rotnitzky, and Zhao (1994), Robins, Rotnitzky, and Zhao (1995), and Bang and Robins (2005) discuss doubly robust estimators of the outcome that are consistent under MAR when either the conditional outcome or the attrition model are correctly specified. This approach satisfies Neyman orthogonality as required for double machine learning.[1] However, their framework does not consider double selection into treatment and the observability of the outcome at the same time as we do in this paper.

Negi (2020) suggests an alternative estimator under double selection that falls into the weighted M-estimation framework described in Sloczyński and Wooldridge (2018) and also satisfies doubly robustness, i.e. remains consistent under parametric misspecification of either the conditional outcome model or the treatment and selection models. This approach based on reweighting outcome models is nevertheless different to ours making use of efficient influence functions and to the best of our knowledge, Neyman (1959) orthogonality (as required for double machine learning) has not been shown for weighted M-estimation (while we prove this property for our proposed estimators). A further difference is that Negi (2020) focuses on treatment evaluation when controlling for pre-treatment covariates to tackle double selection, while we in addition consider identification based on both pre- and post-treatment covariates (dynamic confounding) or an instrument for sample selection.

In contrast to MAR-based identification, so-called sample selection or nonignorable nonresponse models allow for unobserved confounders of the attrition process and the outcome. Unless strong functional form assumptions as in Heckman (1976), Heckman (1979), Hausman and Wise (1979), and Little (1995) are imposed, identification requires an instrumental variable (IV) for sample selection. We refer to Das, Newey, and Vella (2003), Newey (2007), Huber (2012), and Huber (2014b) for nonparametric estimation approaches in this context. To the best of our knowledge, this study is the first one to propose a doubly robust treatment effect estimator under nonignorable outcome attrition and to consider machine learning techniques to control

[1] Relatedly, Barnwell and Chaudhuri (2020) consider several outcome periods under a monotonic MAR assumption (i.e. outcome attrition being an absorbing state weakly increasing over time) and also discuss the evaluation of randomly assigned treatments in this context based on the efficient influence function. In contrast, our framework considers a single outcome period and permits selection into treatment to be related to observed confounders.

for (possibly high-dimensional) covariates in this context. Our estimators are available in the *causalweight* package for R by Bodory and Huber (2018).

This paper proceeds as follows. Using the potential outcome framework, Section 2 discusses the identification of the average treatment effect when outcomes are assumed to be missing at random (i.e. selection is on observables, as for the treatment) conditional on pre-treatment covariates. Section 3 considers identification when outcome attrition is related to unobservables, known as nonignorable nonresponse, and an instrument is available for tackling this issue. Section 4 demonstrates identification under a sequential selection-on-observables which allows for dynamic confounding, meaning that outcomes are assumed to be missing at random conditional on pre- and post-treatment covariates. Section 5 proposes an estimator based on double machine learning and shows root-n consistency and asymptotic normality under specific regularity conditions. Section 6 provides a simulation study. Section 7 presents an empirical application to data from the US Job Corps Study. Section 8 concludes.

2 Identification under missingness at random

Our target parameter is the average treatment effect (ATE) of a binary or multiply discretely distributed treatment variable D on an outcome variable Y. To define the effect of interest, we use the potential outcome framework, see Rubin (1974). Let $Y(d)$ denote the potential outcome under hypothetical treatment assignment $d \in \{0, 1, ..., Q\}$, with 0 indicating non-treatment and $1, ..., Q$ the different treatment choices (where Q denotes the number of non-zero treatments). The ATE when comparing two distinct treatment $d \neq d'$ then corresponds to $\Delta = E[Y(d) - Y(d')]$. Furthermore, let Y denote the outcome realized under the treatment (f)actually assigned to a subject, i.e. $Y = Y(D)$. Therefore, Y corresponds to the potential outcome under the treatment received, while the potential outcome under any counterfactual treatment assignment remains unknown. A further complication in our evaluation framework is that Y is assumed to be only observed for a subpopulation, i.e. conditional on $S = 1$, where S is a binary variable indicating whether Y is observed/selected, or not.

Empirical examples with partially observed outcomes include wage regressions, with S being an employment indicator, see for instance Gronau (1974), or the evaluation of the effects of policy interventions in education on test scores, with S being participation in the test, see

Angrist, Bettinger, and Kremer (2006). Throughout our discussion, S is permitted to be a function of D and X, i.e. $S = S(D, X)$. However, S must neither be affected by nor affect Y.[2] Therefore, selection per se does not causally influence the outcome. The following nonparametric outcome and selection models satisfy this framework:

$$Y = \phi(D, X, U), \quad S = \psi(D, X, V), \tag{1}$$

where U, V are unobserved characteristics and ϕ, ψ are general functions.[3] Throughout the paper we assume that the stable unit treatment value assumption (SUTVA, Rubin (1980)) holds such that $\Pr(D = d \implies Y = Y(d)) = 1$ This rules out interaction or general equilibrium effects and implies that the treatment is uniquely defined.

We subsequently formalize the assumptions that permit identifying the average treatment effect when both selection into the treatment and outcome attrition is related to observed characteristics.

Assumption 1 (conditional independence of the treatment):

$Y(d) \perp D | X = x$ for all $d \in \{0, 1, ..., Q\}$ and x in the support of X.

By Assumption 1, there are no unobservables jointly affecting the treatment and the outcome conditional on covariates X. For model (1), this implies that U is not associated with unobserved terms affecting D given X. In observational studies, the plausibility of this assumption crucially hinges on the richness of the data, while in experiments, it is satisfied if the treatment is randomized within strata defined by X or randomized independently of X.

Assumption 2 (conditional independence of selection):

$Y \perp S | D = d, X = x$ for all $d \in \{0, 1, ..., Q\}$ and x in the support of X.

By Assumption 2, there are no unobservables jointly affecting selection and the outcome conditional on D, X, such that outcomes are missing at random (MAR) in the denomination of Rubin (1976). Put differently, selection is assumed to be selective w.r.t. observed characteristics only. For model (1), this implies that U and V are conditionally independent given D, X.

Assumption 3 (common support):

[2] See for instance Imai (2009) for alternative assumptions, which imply that selection is associated with the outcome but is independent of the treatment conditional on the outcome and other observable variables.

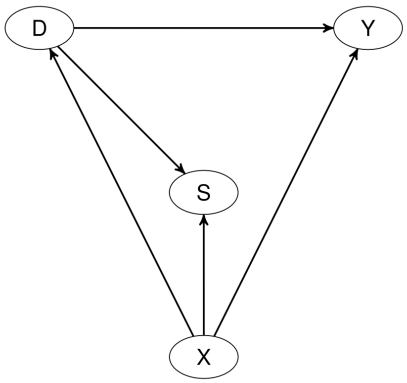
[3] Note that $Y(d) = \phi(d, X, U)$, which means that fixing the treatment yields the potential outcome.

(a) $\Pr(D = d|X = x) > 0$ and (b) $\Pr(S = 1|D = d, X = x) > 0$ for all $d \in \{0, 1, ..., Q\}$ and x in the support of X.

Assumption 3(a) is a common support restriction requiring that the conditional probability to receive a specific treatment given X, henceforth referred to as treatment propensity score, is larger than zero in either treatment state. Assumption 3(b) requires that for any combination of D, X, the conditional probability to be observed, henceforth referred to as selection propensity score, is larger than zero. Otherwise, the outcome is not observed for some specific combinations of these variables implying yet another common support issue.

Figure 1 provides a graphical illustration of our identification setup using a directed acyclic graph, with arrows representing causal effects. Each of D, S, and Y might be causally affected by distinct and statistically independent sets of unobservables not displayed in Figure 1, but none of these unobservables may jointly affect D and Y given X or S and Y given D and X.

Figure 1: Causal paths under the missing at random assumption



Our identifying assumptions imply that

$$E[Y(d)|X] = E[Y|D = d, X] = E[Y|D = d, S = 1, X], \tag{2}$$

where the first equality follows from Assumption 1 and the second equality from Assumption 2.

Therefore, the mean potential outcome identified by

$$E[Y(d)] = E[E[Y|D=d, S=1, X]], \tag{3}$$

or, using the fact that $E[Y|D=d,S=1,X] = E[I\{D=d\}\cdot S\cdot Y|X]/\Pr(D=d,S=1|X)$, by

$$E[Y(d)] = E\left[\frac{E[Y\cdot I\{D=d\}\cdot S|X]}{\Pr(D=d,S=1|X)}\right] = E\left[\frac{I\{D=d\}\cdot S\cdot Y}{\Pr(D=d|X)\cdot\Pr(S=1|D=d,X)}\right], \tag{4}$$

where the second equality follows from the law of iterated expectations. $I\{\cdot\}$ denotes the indicator function, which is equal to one if its argument is satisfied and zero otherwise. Division by $\Pr(D=d|X)\cdot\Pr(S=1|D=d,X)$ in (4) also demonstrates the importance of Assumption 3 for nonparametric identification. For the sake of brevity, we henceforth denote by $\mu(D,S,X) = E[Y|D,S,X]$ the conditional mean outcome and by $p^d(X) = \Pr(D=d|X)$ and $\pi(D,X) = \Pr(S=1|D,X)$ the propensity scores. Expressions (3) and (4) suggest that the mean potential outcomes (and thus, the ATE) are identified, either based on conditional mean outcomes or inverse probability weighting using the treatment and selection propensity scores.

Following the literature on doubly robust methods, see e.g. Robins, Mark, and Newey (1992), Robins, Rotnitzky, and Zhao (1994), and Robins, Rotnitzky, and Zhao (1995), we combine both approaches to obtain the following identification result:

$$\begin{aligned} E[Y(d)] &= E\Big[\psi_d\Big], \text{ where} \\ \psi_d &= \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)} + \mu(d,1,X). \end{aligned} \tag{5}$$

The result in (5) is based on the so-called efficient score function, which is formally derived in Appendix B following the approach outlined in Levy (2019). By noting that

$$\begin{aligned} & E\left[\frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)}\right] = E\left[\frac{E[I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]|X]}{p_d(X)\cdot\pi(d,X)}\right] \\ = \;& E[E[Y-\mu(d,1,X)|D=d,S=1,X]] = E[E[Y|D=d,S=1,X]-\mu(d,1,X)] \\ = \;& E[\mu(d,1,X)-\mu(d,1,X)] = 0, \end{aligned} \tag{6}$$

it is easy to see that (5) is equivalent to (3) and thus, (4). In contrast to (3) and (4), however, expression (5) is doubly robust in the sense that it identifies $E[Y(d)]$ if either the conditional

mean outcome $\mu(d,1,X)$ or the propensity scores $p_d(X)$ and $\pi(d,X)$ are correctly specified. Furthermore, it satisfies the so-called Neyman (1959) orthogonality, i.e. is first-order insensitive to perturbations in $\mu(D,S,X)$, $p_d(X)$, and $\pi(D,X)$, see Appendix A.1. This entails desirable robustness properties when using machine learning to estimate the outcome, treatment, and selection models in a data-driven way.

3 Identification under nonignorable nonresponse

When sample selection or outcome attrition is related to unobservables even conditional on observables, identification generally requires an instrument for S. We therefore replace Assumptions 2 and 3, but maintain Assumption 1 (i.e. selection into treatment is on observables).

Assumption 4 (Instrument for selection):

(a) There exists an instrument Z that may be a function of D, i.e. $Z = Z(D)$, is conditionally correlated with S, i.e. $E[Z \cdot S|D,X] \neq 0$, and satisfies (i) $Y(d,z) = Y(d)$ and (ii) $Y \perp Z|D = d, X = x$ for all $d \in \{0,1,...,Q\}$ and x in the support of X,

(b) $S = I\{V \leq \chi(D,X,Z)\}$, where χ is a general function and V is a scalar (index of) unobservable(s) with a strictly monotonic cumulative distribution function conditional on X,

(c) $V \perp (D,Z)|X$.

Assumption 4 no longer imposes the conditional independence of Y and S given D,X. As the unobservable V in the selection equation is allowed to be associated with unobservables affecting the outcome, Assumptions 1 and 2 generally do not hold conditional on $S = 1$ due to the endogeneity of the post-treatment variable S. In fact, $S = 1$ implies that $\chi(D,X,Z) > V$ such that conditional on X, the distribution of V generally differs across values of D. This entails a violation of the conditional independence of D and $Y(d)$ given $S = 1$ and X if the potential outcome distributions differ across values of V. We therefore require an instrumental variable denoted by Z, which must not affect Y or be associated with unobservables affecting Y conditional on D and X, as invoked in 4(a).[4] We apply a control function approach based on this instrument,[5] which requires further assumptions.

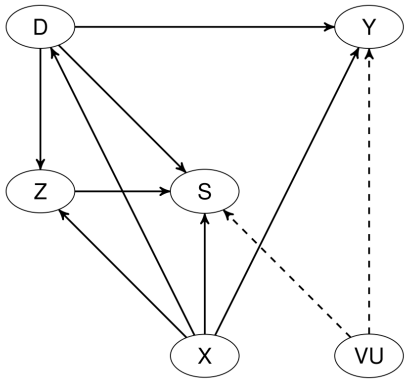
[4] As an alternative set of IV restrictions in the context of selection, d'Haultfoeuille (2010) permits the instrument to be associated with the outcome, but assumes conditional independence of the instrument and selection given the outcome.

[5] Control function approaches have been applied in semi- and nonparametric sample selection models, e.g. Ahn and Powell (1993), Das, Newey, and Vella (2003), Newey (2007), Huber (2012), and Huber (2014b), as well as

By the threshold crossing model postulated in 4(b), $\Pr(S = 1|D, X, Z) = \Pr(V \leq \chi(D, X, Z)) = F_V(\chi(D, X, Z))$, where $F_V(v)$ denotes the cumulative distribution function of V evaluated at v. We will henceforth use the notation $\Pi = \pi(D, X, Z) = \Pr(S = 1|D, X, Z)$ for the sake of brevity. Again by Assumption 4(b), the selection probability Π increases strictly monotonically in χ, such that there is a one-to-one correspondence between the distribution function F_V and specific values v given X. By Assumption 4(c), V is independent of (D, Z) given X, implying that the distribution function of V given X is (nonparametrically) identified. By comparing individuals with the same Π, we control for F_V and thus for the confounding associations of V with D and $Y(d)$ that occur conditional on $S = 1, X$. In other words, Π serves as control function where the exogenous variation comes from Z. Controlling for the distribution of V based on the instrument is thus a feasible alternative to the (infeasible) approach of directly controlling for levels of V.

Figure 2 provides an acyclic graph of a causal model that can satisfy Assumptions 1 and 4. U denotes unobservables affecting the outcome, which may be arbitrarily associated with V, the unobservable affecting selection. Note that the dashed lines indicate that V, U are not observed in the data. Identification relies on instrument Z, which must not be associated with outcome Y conditional on D and X.

Figure 2: Causal paths under nonignorable nonresponse



in nonparametric instrumental variable models, see for example Newey, Powell, and Vella (1999), Blundell and Powell (2004), and Imbens and Newey (2009).

Furthermore, identification requires the following common support assumption, which is similar to Assumption 3(a), but in contrast to the latter also includes Π as a conditioning variable.

Assumption 5 (common support):

$\Pr(D = d|X = x, \Pi = \pi) > 0$ for all $d \in \{0, 1, ..., Q\}$ and x, z in the support of X, Z.

This means that in fully nonparametric contexts, the instrument Z must in general be continuous and strong enough to importantly shift the selection probability Π conditional on D, M, X in the selected population. Assumptions 1, 4, and 5 are sufficient for the identification of mean potential outcomes and the ATE in the selected population, denoted as $\Delta_{S=1} = E[Y(1) - Y(0)|S = 1]$.

To see this, note that the identifying assumptions imply

$$E[Y(d)|S = 1, X, F_V] = E[Y(d)|S = 1, X, \Pi] = E[Y|D = d, S = 1, X, \Pi] \tag{7}$$

The first equality follows from $\Pi = F_V$ under Assumption 4, the second from the fact that when controlling for F_V, conditioning on $S = 1$ does not result in an association between $Y(d)$ and D given X such that $Y(d) \perp D|X, \Pi, S = 1$ holds by Assumptions 1 and 4. Therefore

$$E[Y(d)|S = 1] = E[E[Y|D = d, S = 1, X, \Pi]|S = 1]. \tag{8}$$

Denoting by $p_d(X, \Pi) = \Pr(D = d|X, \Pi)$ and $\mu(D, S, X, \Pi) = E[Y|D, S, X, \pi(D, X, Z)]$, an alternative expression for the mean potential outcome among the selected is obtained by

$$\begin{aligned} E[Y(d)|S = 1] &= E\Big[\phi_{d,S=1}|S = 1\Big], \text{ where} \\ \phi_{d,S=1} &= \frac{I\{D = d\} \cdot [Y - \mu(d, 1, X, \Pi)]}{p_d(X, \Pi)} + \mu(d, 1, X, \Pi), \end{aligned} \tag{9}$$

where division by $p_d(X, \Pi)$ makes the reliance on Assumption 5 explicit. By applying the law of iterated expectations to replace $[Y - \mu(d, 1, X, \Pi)]$ with $E[Y - \mu(d, 1, X, \Pi)|D = d, S = 1, X, \Pi]$ and noting that the latter expression is zero, one can see that (9) is equivalent to (8). But in contrast to the latter, the identification result in (9) satisfies Neyman orthogonality and is based on the efficient influence function,[6] see Appendix B.

[6] While the efficient influence function associated with (9) is technically speaking doubly robust, i.e. consistent if either $\mu(d, 1, X, \Pi)$ or $p_d(X, \Pi)$ is correctly specified, it is worth noting that this property can generally only hold

The identification of the ATE in the total (rather than the selected) population is not feasible without further assumptions. The reason is that effects among selected observations cannot be extrapolated to the non-selected population if the effect of D interacts with unobservables affecting the outcome, i.e. U in (1), as the latter are in general distributed differently across $S = 1, 0$ even conditional on (X, Π) or (D, X, Π). To see this, note that conditional on $\Pi = \Pr(V \leq \chi(D, X, Z))$, the distribution of V differs across the selected (satisfying $V \leq \chi(D, X, Z)$) and the non-selected (satisfying $V > \chi(D, X, Z)$), such that the distribution of U differs, too, if V and U are associated. This generally implies that $E[Y(1) - Y(0)|S = 1, X, \Pi] \neq E[Y(1) - Y(0)|S = 0, X, \Pi]$. While control function Π ensures (together with X) that the treatment is unconfounded in the selected subpopulation, it does not permit extrapolating effects to the non-selected population with unobserved outcomes, see also Huber and Melly (2015) for further discussion.

Assumption 6 therefore imposes homogeneity in the average treatment effect across selected and non-selected populations conditional on X, V. A sufficient condition for effect homogeneity is the separability of observed and unobserved components in the outcome equation, i.e. $Y = \eta(D, X) + \nu(U)$, where η, ν are general functions. Furthermore, common support as postulated in Assumption 5 needs to be strengthened to hold in the entire population. In addition, the selection probability Π must be larger than zero for any d, x, z in their support. Otherwise, outcomes are not observed for some values of D, X. Assumption 7 formalizes this common support restriction.

Assumption 6 (conditional effect homogeneity):

$E[Y(d) - Y(d')|S = 1, X = x, V = v] = E[Y(d) - Y(d')|X = x, V = v]$ for all $d \neq d' \in \{0, 1, ..., Q\}$ and x, v in the support of X, V.

Assumption 7 (common support):

$\pi(d, x, z) > 0$ for all $d \in \{0, 1, ..., Q\}$ and x, z in the support of X, Z.

---

if Π is correctly specified because it enters both $\mu(d, 1, X, \Pi)$ and $p_d(X, \Pi)$ as first step estimator. However, our approach does not rely on (global) doubly robustness but on Neyman orthogonality, which implies that DML is robust to local perturbations in Π under specific regularity conditions.

Under Assumptions 1,4,5,6, and 7, it follows that

$$\mu(d,1,X,\Pi) - \mu(d',1,X,\Pi) = E[Y(d) - Y(d')|S=1,X,V] = E[Y(d) - Y(d')|X,V], \quad (10)$$

where the first equality follows from Assumptions 1 and 4, see (7), and the second one from Assumption 6. Therefore, the ATE is identified by

$$\Delta = E[\mu(d,1,X,\Pi) - \mu(d',1,X,\Pi)]. \quad (11)$$

An alternative expression for the ATE that is based on the efficient influence function and respects Neyman orthogonality is given by

$$\begin{aligned} \Delta &= E\Big[\phi_d - \phi_{d'}\Big], \text{ where} \\ \phi_d &= \frac{I\{D=d\} \cdot S \cdot [Y - \mu(d,1,X,\Pi)]}{p_d(X,\Pi) \cdot \pi(d,X,Z)} + \mu(d,1,X,\Pi), \end{aligned} \quad (12)$$

where division by $p_d(X,\Pi) \cdot \pi(d,X,Z)$ relies on the satisfaction of Assumptions 5 and 7.

4 Identification under sequential conditional independence

In many applications, it might appear unrealistic that one can control for all variables jointly affecting the sample selection indicator by conditioning only on baseline covariates measured prior to treatment assignment, in particular when no instrument is at hand. This is particularly the case when there is a substantial time lag between treatment assignment and sample selection/attrition, which raises concerns about dynamic confounding. The latter implies that some confounders influencing both the outcome and sample selection are themselves a function of the treatment. We subsequently reconsider the MAR framework, but not modify the identifying assumptions such that observed post-treatment confounders of Y and S are permitted. We will subsequently refer to observed post-treatment variables by M, in order to distinguish them from pre-treatment covariates X. Identification is based on a sequential conditional independence, which is based on maintaining Assumption 1 (conditional independence of D given X), but replacing Assumption 2 by a modified conditional independence assumption for the selection indicator S that allows for dynamic confounding due to $M = M(D)$, i.e. covariates possibly

influenced by the treatment.

Assumption 8 (conditional independence of selection):

$Y \perp S|D = d, X = x, M = m$ for all $d \in \{0, 1, ..., Q\}$ and x, m in the support of X and M.

By Assumption 8, there are no unobservables jointly affecting selection and the outcome conditional on D, X, M, such that sample selection is selective w.r.t. observed characteristics only. When modifying the nonparametric outcome and selection models in (1) to $Y = \phi(D, X, M, U)$ and $S = \psi(D, X, M, V)$, Assumption 8 is satisfied if unobservables U and V are independent.

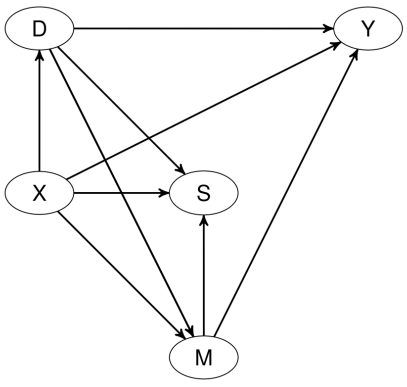
Assumption 9 (common support):

(a) $\Pr(D = d|X = x) > 0$ and (b) $\Pr(S = 1|D = d, X = x, M = m) > 0$ for all $d \in \{0, 1, ..., Q\}$ and x, m in the support of X, M.

Part (b) in Assumption 9 imposes a somewhat stronger common support restriction than part (b) in Assumption 3, as it requires the selection propensity score to be larger than zero for any combination of D, X, M (rather D, X only).

Figure 3 provides an acyclic graph in which Assumptions 1 and 8 hold. Post-treatment covariates M may be influenced by D, X and might jointly affect S and Y. Conditional on D, X, M, there are, however, no unobservables jointly influencing S and Y.

Figure 3: Causal paths under sequential conditional independence



Our identifying assumptions imply that

$$\begin{aligned} E[Y(d)] &= E[E[Y(d)|X]] = E[E[Y|D=d,X]] = E[E[E[Y|D=d,X,M]|D=d,X]] \\ &= E[E[E[Y|D=d,S=1,X,M]|D=d,X]]. \end{aligned} \tag{13}$$

where the first and third equalities follow from the law of iterated expectations, the second from Assumption 1, and the fourth from Assumption 8. Alternatively to this regression-based result using nested conditional mean outcomes, an IPW-based expression can be obtained, in which we use $\pi(D,X,M) = \Pr(S=1|D,X,M)$ as shortcut notation for the selection propensity score.

$$\begin{aligned} & E[E[Y|D=d,S=1,X,M|D=d,X]] = E\left[E\left[E\left[\frac{S\cdot Y}{\pi(d,X,M)}\middle|D=d,X,M\right]\middle|D=d,X\right]\right] \\ = \; & E\left[E\left[\frac{S\cdot Y}{\pi(d,X,M)}\middle|D=d,X\right]\right] = E\left[E\left[\frac{I\{D=d\}\cdot S\cdot Y}{p_{d0}(X)\cdot\pi_0(d,X,M)}\middle|X\right]\right] \\ = \; & E\left[\frac{I\{D=d\}\cdot S\cdot Y}{p_{d0}(X)\cdot\pi_0(d,X,M)}\right], \end{aligned} \tag{14}$$

where the first and third equalities follow from basic probability theory and the second and last ones from the law of iterated expectations. Combining regression and IPW yields the following doubly robust identification result based on the efficient influence function, in which $\mu(d,1,X,M) = E[Y|D=d,S=1,X,M]$ and $\nu(d,1,X) = E[E[Y|D=d,S=1,X,M]|D=d,X]$ denote the conditional mean outcome and the nested conditional mean outcome, respectively:

$$\begin{aligned} E[Y(d)] &= E\Big[\theta_d\Big], \text{ where} \\ \theta_d &= \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X,M)]}{p_d(X)\cdot\pi(d,X,M)} \\ &+ \frac{I\{D=d\}\cdot[\mu(d,1,X,M)-\nu(d,1,X)]}{p_d(X)} + \nu(d,1,X), \end{aligned} \tag{15}$$

where division by $p_d(X)\cdot\pi(d,X,M)$ relies on Assumption 9. The derivation of the efficient influence function is provided in Appendix B.

5 Estimation of the counterfactual with K-fold Cross-Fitting

We subsequently propose an estimation strategy for the counterfactual $E[Y(d)]$ under MAR as discussed in Section 2 based on identification result (5) and show its root-n consistency under specific regularity conditions. Let to this end $\mathcal{W} = \{W_i | 1 \leq i \leq n\}$ with $W_i = (Y_i \cdot S_i, D_i, S_i, X_i)$ for all i denote the set of observations in an i.i.d. sample of size n. η denotes the plug-in (or nuisance) parameters, i.e. the conditional mean outcome, mediator density and treatment probability. Their respective estimates are referred to by $\hat{\eta} = \{\hat{\mu}(D, 1, X), \hat{p}_d(X), \hat{\pi}(D, X)\}$ and the true parameters by $\eta_0 = \{\mu_0(D, 1, X), p_{d0}(X), \pi_0(D, X)\}$. Finally, $\Psi_{d0} = E[Y(d)]$ denotes the true counterfactual.

We estimate Ψ_{d0} by the following algorithm that combines the estimation of Neyman-orthogonal scores with sample splitting or cross-fitting and is root-n consistent under conditions outlined further below.

Algorithm 1: Estimation of $E[Y(d)]$ based on equation (5)

1. Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in k.

2. For each k, use $\mathcal{W}_k^C$ to estimate the model parameters of the plug-ins $\mu(D, S = 1, X)$, $p_d(X)$, $\pi(D, X)$ in order to predict these plug-ins in $\mathcal{W}_k$, where the predictions are denoted by $\hat{\mu}^k(D, 1, X)$, $\hat{p}_d^k(X)$, and $\hat{\pi}^k(D, X)$.

3. For each k, obtain an estimate of the score function (see ψ_d in (5)) for each observation i in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}^k$:

$$\hat{\psi}_{d,i}^k = \frac{I\{D_i = d\} \cdot S_i \cdot [Y_i - \hat{\mu}^k(d, 1, X_i)]}{\hat{p}_d^k(X_i) \cdot \hat{\pi}^k(d, X_i)} + \hat{\mu}^k(d, 1, X_i). \tag{16}$$

4. Average the estimated scores $\hat{\psi}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\Psi_{d0} = E[Y(d)]$ in the total sample, denoted by $\hat{\Psi}_d = 1/n \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$.

In order to obtain root-n consistency for counterfactual estimation, we make the following assumption about the prediction qualities of machine learning for estimating the nuisance parameters. Following Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins

(2018), we introduce some further notation: let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ denote sequences of positive constants with $\lim_{N\to\infty}\delta_n = 0$ and $\lim_{N\to\infty}\Delta_n = 0$. Furthermore, let c, ϵ, C and q be positive constants such that $q > 2$, and let $K \geq 2$ be a fixed integer. Also, for any random vector $R = (R_1, ..., R_l)$, let $\|R\|_q = \max_{1\leq j\leq l}\|R_l\|_q$, where $\|R_l\|_q = (E[|R_l|^q])^{\frac{1}{q}}$. In order to ease notation, we assume that n/K is an integer. For the sake of brevity we omit the dependence of probability $\Pr_P$, expectation $E_P(\cdot)$, and norm $\|\cdot\|_{P,q}$ on the probability measure P.

Assumption 10 (regularity conditions and quality of plug-in parameter estimates): For all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws the following conditions hold for the random vector (Y, D, S, X) for $d \in \{0, 1, ..., Q\}$:

(a) $\|Y\|_q \leq C$,

$\left\|E[Y^2|D = d, S = 1, X]\right\|_\infty \leq C^2$,

(b) $\Pr(\epsilon \leq p_{d0}(X) \leq 1 - \epsilon) = 1$,

$\Pr(\epsilon \leq \pi_0(d, X)) = 1$,

(c) $\|Y - \mu_0(d, 1, X)\|_2 = E\Big[\,(Y - \mu_0(d, 1, X))^2\,\Big]^{\frac{1}{2}} \geq c$

(d) Given a random subset I of $[n]$ of size $n_k = n/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i\in I^C})$ satisfies the following conditions. With P-probability no less than $1 - \Delta_n$:

$$\begin{aligned}
\|\hat{\eta}_0 - \eta_0\|_q &\leq C,\\
\|\hat{\eta}_0 - \eta_0\|_2 &\leq \delta_n,\\
\|\hat{p}_{d0}(X) - 1/2\|_\infty &\leq 1/2 - \epsilon,\\
\|\hat{\pi}_0(D, X) - 1/2\|_\infty &\leq 1/2 - \epsilon,\\
\|\hat{\mu}_0(D, S, X) - \mu_0(D, S, X)\|_2 \times \|\hat{p}_{d0}(X) - p_0(X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\hat{\mu}_0(D, S, X) - \mu_0(D, S, X)\|_2 \times \|\hat{\pi}_0(D, X) - \pi_0(D, X)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}$$

The only non-primitive condition is the condition (d), which puts restrictions on the quality of the nuisance parameter estimators. Condition (a) states that the distribution of the outcome

does not have unbounded moments. (b) refines the common support condition such that the treatment and selection propensity scores are bounded away from 0and 1 and 0, respectively. (c) states that covariates X do not perfectly predict the conditional mean outcome.

For demonstrating the root-n consistency of our estimator of the mean potential outcome, we show that it satisfies the requirements of the DML framework in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) by first verifying linearity and Neyman orthogonality of the score (see Appendix A.1). As $\psi_d(W, \eta, \Psi_{d0})$ is smooth in (η, Ψ_{d0}), it then suffices that the plug-in estimators converge with rate $n^{-1/4}$ for achieving $n^{-1/2}$-convergence in the estimation of $\hat{\psi}$, see Theorem 1. A rate of $n^{-1/4}$ is achievable by many commonly used machine learners under specific conditions, such as lasso, random forests, boosting and neural nets, see for instance Belloni, Chernozhukov, and Hansen (2014), Luo and Spindler (2016), Wager and Athey (2018), and Farrell, Liang, and Misra (2018).

Theorem 1

Under Assumptions 1-3 and 10, it holds for estimating $\Psi_{d0} = E[Y(d)]$ based on Algorithm 1: $\sqrt{n}\left(\hat{\Psi}_d - \Psi_{d0}\right) \to N(0, \sigma^2_{\psi_d})$, where $\sigma^2_{\psi_d} = E[(\psi_d - \Psi_{d0})^2]$.
The proof is provided in Appendix A.1.

We subsequently discuss the estimation of Ψ_{d0} based on (12). We note that in this case, one needs to estimate the nested nuisance parameters $\mu(d, 1, X, \Pi)$ and $p_d(X, \Pi)$, because they require the first-step estimation of $\Pi = \pi(D, X, Z)$. To avoid overfitting in the nested estimation procedure, the models for Π on the one hand and $\mu(d, 1, X, \Pi), p_d(X, \Pi)$ on the other hand are estimated in different subsamples. The plug-in estimates are now denoted by $\hat{\eta} = \{\hat{\mu}(D, 1, X, \Pi), \hat{p}_d(X, \Pi), \hat{\pi}(D, X, Z)\}$ and the true plug-ins by $\eta_0 = \{\mu_0(D, 1, X, \Pi), p_{d0}(X, \Pi), \pi_0(D, X, Z)\}$.

Algorithm 2: Estimation of $E[Y(d)]$ based on equation (12)

1. Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in k.

2. Split $\mathcal{W}_k^C$ into 2 nonoverlapping subsamples and estimate the model parameters of $\pi_0(D, X, Z)$ in one subsample and the model parameters of $\mu_0(D, 1, X, \Pi)$ and $p_{d0}(X, \Pi)$ in the other subsample. Predict the plug-in models in $\mathcal{W}_k$, where the predictions are denoted by $\hat{\Pi}^k$, $\hat{p}_d^k(X, \hat{\Pi}^k)$, and $\hat{\mu}(D, 1, X, \hat{\Pi})$.

3. For each k, obtain an estimate of the efficient score function (see ϕ_d in (12)) for each observation i in $\mathcal{W}_k$, denoted by $\hat{\phi}^k_{d,i}$:

$$\hat{\phi}^k_{d,i} = \frac{I\{D_i = d\} \cdot S_i \cdot [Y_i - \hat{\mu}^k(d, 1, X_i, \hat{\Pi}_i)]}{\hat{p}_d(X_i, \hat{\Pi}_i) \cdot \hat{\pi}(d, X_i, Z_i)} + \hat{\mu}(d, 1, X_i, \hat{\Pi}_i) \tag{17}$$

4. Average the estimated scores $\hat{\phi}^k_{d,i}$ over all observations across all K subsamples to obtain an estimate of $\Psi_{d0} = E[Y(d)]$ in the total sample, denoted by $\hat{\Phi}_d = 1/n \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\phi}^k_{d,i}$.

An estimator of $\Psi^{S=1}_{d0} = E[Y(d)|S = 1]$ based on (9) is obtained by two modifications in Algorithm 2. First, rather than relying on the total sample n, one merely uses the subsample with observed outcomes which of size $\sum_{i=1}^{n} S_i$ to split it into K subsamples. Second, in step 3, $\hat{\phi}^k_{d,i}$ is to be replaced by

$$\hat{\phi}^k_{d,S=1,i} = \frac{I\{D_i = d\} \cdot [Y_i - \hat{\mu}^k(d, 1, X_i, \hat{\Pi}_i)]}{\hat{p}_d(X_i, \hat{\Pi}_i)} + \hat{\mu}(d, 1, X_i, \hat{\Pi}_i) \tag{18}$$

to estimate $\Psi^{S=1}_{d0}$ in step 4 by $\hat{\Phi}^{S=1}_d = \frac{1}{\sum_{i=1}^{n} S_i} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\phi}^k_{d,S=1,i}$. As $\sum_{i=1}^{n} S_i$ is an asymptotically fixed proportion of n, also this approach can be shown to be root-n consistent under particular regularity conditions outlined in Assumption 11, that are in analogy to those in Assumption 10, but now adapted our IV-dependent identifying assumptions.

Assumption 11 (regularity conditions and quality of plug-in parameter estimates):

For all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws the following conditions hold for the random vector (Y, D, S, X, Z) for $d \in \{0, 1, ..., Q\}$:

(a) $\|Y\|_q \leq C$,

$\left\|E[Y^2|D = d, S = 1, X, \Pi]\right\|_\infty \leq C^2$,

(b) $\Pr(\epsilon \leq p_{d0}(X, \Pi) \leq 1 - \epsilon) = 1$,

$\Pr(\epsilon \leq \pi_0(d, X, Z)) = 1$,

(c) $\|Y - \mu_0(d, 1, X, \Pi)\|_2 = E\Big[(Y - \mu_0(d, 1, X, \Pi))^2 \Big]^{\frac{1}{2}} \geq c$

(d) Given a random subset I of $[n]$ of size $n_k = n/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^C})$ satisfies the following conditions. With P-probability no less than

$1 - \Delta_n$:

$$
\begin{aligned}
\|\hat{\eta}_0 - \eta_0\|_q &\leq C, \\
\|\hat{\eta}_0 - \eta_0\|_2 &\leq \delta_n, \\
\left\|\hat{p}_{d0}(X, \hat{\Pi}) - 1/2\right\|_\infty &\leq 1/2 - \epsilon, \\
\|\hat{\pi}_0(D, X, Z) - 1/2\|_\infty &\leq 1/2 - \epsilon, \\
\left\|\hat{\mu}_0(D, S, X, \hat{\Pi}) - \mu_0(D, S, X, \Pi)\right\|_2 \times \left\|\hat{p}_{d0}(X, \hat{\Pi}) - p_0(X, \Pi)\right\|_2 &\leq \delta_n n^{-1/2}, \\
\left\|\hat{\mu}_0(D, S, X, \hat{\Pi}) - \mu_0(D, S, X, \Pi)\right\|_2 \times \|\hat{\pi}_0(D, X, Z) - \pi_0(D, X, Z)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}
$$

Theorems 2 and 3 postulate the root-n consistency and asymptotic normality of the estimators of the mean potential outcomes in the selected and total populations, respectively.

Theorem 2

Under Assumptions 1, 4, 6, 7, and 11, it holds for estimating $\Psi_{d0} = E[Y(d)]$ based on Algorithm 2:

$\sqrt{n}\left(\hat{\Phi}_d - \Psi_{d0}\right) \to N(0, \sigma^2_{\phi_d})$, where $\sigma^2_{\phi_d} = E[(\phi_d - \Psi_{d0})^2]$.

Theorem 3

Under Assumptions 1, 4, 5, and 11, it holds for estimating $\Psi^{S=1}_{d0} = E[Y(d)|S = 1]$ based on Algorithm 2:

$\sqrt{n}\left(\hat{\Phi}^{S=1}_d - \Psi^{S=1}_{d0}\right) \to N(0, \sigma^2_{\phi_{d,S=1}})$, where $\sigma^2_{\phi_{d,S=1}} = E[(\phi_{d,S=1} - \Psi^{S=1}_{d0})^2]$.

The proofs are provided in Appendices A.2 and A.3.

Next, we consider the estimation of Ψ_{d0} based on (15). Similarly to estimation based on (12), we are required to estimate a nested nuisance parameter, namely $\nu(d, 1, X) = E[\mu(d, 1, X, M)|D = d, X]$. To avoid overfitting in the nested estimation procedure, the models for $\mu(d, 1, X, M)$ and $\nu(d, 1, X)$ estimated in different subsamples.

Algorithm 3: Estimation of $E[Y(d)]$ based on equation (15)

1. Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}^C_k$ the complement set of all observations not in k.

2. For each k, use $\mathcal{W}^C_k$ to estimate the model parameters of $p_d(X)$ and $\pi(d, X, M)$. Split $\mathcal{W}^C_k$ into 2 nonoverlapping subsamples and estimate the model parameters of the conditional

mean $\mu(d,1,X,M)$ and the nested conditional mean $\nu(d,1,X)$ in the distinct subsamples. Predict the models among $\mathcal{W}_k$, where the predictions are denoted by $\hat{p}_d^k(X)$, $\hat{\pi}^k(d,X,M)$, $\hat{\mu}^k(d,1,X,M)$, $\hat{\nu}^k(d,1,X)$.

3. For each k, obtain an estimate of the moment condition for each observation i in $\mathcal{W}_k$, denoted by $\hat{\theta}_{d,i}^k$:

$$\begin{aligned}\hat{\theta}_{d,i}^k &= \frac{I\{D_i=d\}\cdot S_i\cdot[Y_i-\hat{\mu}^k(d,1,X_i,M_i)]}{\hat{p}_d^k(X_i)\cdot\hat{\pi}^k(d,X_i,M_i)} \\ &+ \frac{I\{D_i=d\}\cdot[\hat{\mu}^k(d,1,X_i,M_i)-\hat{\nu}^k(d,1,X_i)]}{\hat{p}_d^k(X_i)}+\hat{\nu}^k(d,1,X_i).\end{aligned}$$

4. Average the estimated scores $\hat{\theta}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\Psi_{d0}=E[Y(d)]$ in the total sample, denoted by $\hat{\Theta}_d = 1/n\sum_{k=1}^K\sum_{i=1}^{n_k}\hat{\theta}_{d,i}^k$.

To show root-n consistency for this estimation approach, we impose the following regularity conditions, where we again assume that n/K is an integer and omit the dependence of probability $\Pr_P$, expectation $E_P(\cdot)$, and norm $\|\cdot\|_{P,q}$ on the probability measure P:

Assumption 12 (regularity conditions and quality of plug-in parameter estimates):
For all probability laws $P\in\mathcal{P}$ the following conditions hold for the random vector (Y,D,S,X,M) for all $d\in\{0,1,...,Q\}$:

(a) $\|Y\|_q \le C$,

$\left\|E[Y^2|D=d,S=1,X,M]\right\|_\infty \le C^2$,

(b) $\Pr(\epsilon\le p_{d0}(X)\le 1-\epsilon)=1$,

$\Pr(\epsilon\le \pi_0(d,X,M)\le 1-\epsilon)=1$,

(c) $\|Y-\mu_0(d,1,X,M)\|_2 = E\Big[\,(Y-\mu_0(d,1,X,M))^2\,\Big]^{\frac{1}{2}}\ge c$

(d) Given a random subset I of $[n]$ of size $n_k=n/K$, the nuisance parameter estimator $\hat{\eta}_0=\hat{\eta}_0((W_i)_{i\in I^C})$ satisfies the following conditions. With P-probability no less than

$1 - \Delta_n$:

$$\begin{aligned}
\|\hat{\eta}_0 - \eta_0\|_q &\leq C, \\
\|\hat{\eta}_0 - \eta_0\|_2 &\leq \delta_n, \\
\|\hat{p}_{d0}(X) - 1/2\|_\infty &\leq 1/2 - \epsilon, \\
\|\hat{\pi}_0(D, X, M) - 1/2\|_\infty &\leq 1/2 - \epsilon, \\
\|\hat{\mu}_0(D, S, X, M)\|_2 \times \|\hat{p}_{d0}(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}, \\
\|\hat{\mu}_0(D, S, X, M) - \mu_0(D, S, X, M)\|_2 \times \|\hat{\pi}_0(D, X, M) - \pi_0(D, X, M)\|_2 &\leq \delta_n n^{-1/2}, \\
\|\hat{\nu}_0(D, S, X) - \nu_0(D, S, X)\|_2 \times \|\hat{p}_{d0}(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}$$

Under these regularity conditions and the sequential conditional independence assumption, estimation based on Algorithm 3 is asymptotically normal, as postulated in Theorem 4.

Theorem 4

Under Assumptions 1, 8, 9, and 12, it holds for estimating $E[Y(d)]$ based on Algorithm 3:
$\sqrt{n}\Big(\hat{\Theta}_d - \Psi_{d0}\Big) \to N(0, \sigma^2_{\theta_d})$, where $\sigma^2_{\theta_d} = E[(\theta_d - \Psi_{d0})^2]$.
The proof of Theorem 4 is provided in Appendix A.4.

6 Simulation study

This section provides a simulation study to investigate the finite sample behavior of our estimation approaches either relying on a MAR assumption of an instrument for selection based on the following data generating process:

$$\begin{aligned}
Y &= D + X'\beta + U \text{ with } Y \text{ being observed if } S = 1, \\
S &= I\{D + \gamma Z + X'\beta + V > 0\}, \quad D = I\{X'\beta + W > 0\}, \\
X &\sim N(0, \sigma^2_X), \quad Z \sim N(0,1), \quad (U, V) \sim N(0, \sigma^2_{U,V}), \quad W \sim N(0,1).
\end{aligned}$$

Outcome Y is a linear function of D (whose treatment effect is one), covariates X (for $\beta \neq 0$), and the unobservable U and is only observed if the selection indicator S is equal to one. Selection is a function of D, X, the unobservable V, and of instrument Z if $\gamma \neq 0$. The treatment D

is a function of X and the unobservable W. Both Z and W are random, standard normally distributed variables that are uncorrelated with X or (U, V). The correlation between the mean zero and normally distributed covariates in X is determined by the covariance matrix σ_X^2. Similarly, $\sigma_{U,V}^2$ determines the correlation between the mean zero and normally distributed unobservables in the outcome and selection equation. In this setup, MAR is violated if the covariance between U and V is non-zero. We consider the perfromance of our estimators in 1000 simulations with two sample sizes of $n = 2000$ and 8000.

In our simulations, we set the number of covariates p to 100. σ_X^2 is defined based on setting the covariance of the ith and jth covariate in X to $0.5^{|i-j|}$. β gauges the impacts of the covariates on Y, S, and D, respectively, and thus, the magnitude of confounding. The ith element in the coefficient vector β is set to $0.4/i^2$ for $i = 1, ..., p$, implying a squared decay of covariate importance in terms of confounding. In our first simulation design, we set $\gamma = 0$ and $\sigma_{U,V}^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ such that MAR as discussed in Section 2 holds. We consider the performance of DML based on Theorem 1 (henceforth DML MAR), which does not make use of the instrument Z, as well as based on Theorem 2 (DML IV), which exploits the instrument despite the satisfaction of MAR.

The nuisance parameters, i.e. the linear and probit specifications of the outcome, selection, and treatment equations, are estimated by lasso regressions using the default options of the *SuperLearner* package provided by van der Laan, Polley, and Hubbard (2007) for the statistical software R. We use 3-fold cross-fitting for the estimation of the treatment effects. We drop observations whose products of estimated treatment and selection propensity scores are close to zero, namely smaller than a trimming threshold of 0.01 (or 1%). This avoids an explosion of the propensity score-based weights and, thus, of the variance when estimating the mean potential outcomes or ATE by the sample analogues of (5) and (12), where the product of the propensity scores enters the respective denominators for reweighing the outcome. Our estimation procedure is available in the *treatselDML* command of the *causalweight* package for R by Bodory and Huber (2018).

Table 1 presents the simulation results. The biases (bias) of both DML MAR and DML IV are rather close to zero independent of the sample size. Furthermore, the estimators have virtually the same variance, despite the fact that DML IV unnecessarily relies on the control

Table 1: Simulation results under MAR

	true	bias	sd	RMSE	meanSE	coverage
n=2000						
DML MAR	1.000	0.010	0.061	0.062	0.065	0.953
DML IV	1.000	0.010	0.061	0.062	0.065	0.953
n=8000						
DML MAR	1.000	0.014	0.033	0.036	0.035	0.932
DML IV	1.000	0.014	0.033	0.036	0.035	0.932

Notes: column 'true' shows the true effect, 'bias' the bias of the respective estimator, 'sd' the standard deviation, and 'RMSE' the root mean squared error. Column 'meanSE' displays the average standard error based on the asymptotic approximation across all simulations, 'coverage' the coverage rate of the true effect based on 95% confidence intervals.

function approach and an irrelevant instrument. Both estimators appear to converge to the true effect with $\sqrt{n}$-rate, as the root mean squared error (RMSE) is roughly cut by half when quadrupling the sample size. The average standard error across simulations (meanSE) based on the asymptotic variance approximation comes close to the respective estimator's standard deviation (sd). Finally, the coverage rate (coverage), i.e. the share of simulations in which the 95% confidence interval includes the true effect, is close to the nominal level of 95%.

Table 2: Simulation results under nonignorable selection

	true	bias	sd	RMSE	meanSE	coverage
n=2000						
DML MAR	1.000	-0.120	0.056	0.133	0.058	0.462
DML IV	1.000	-0.025	0.074	0.078	0.078	0.933
n=8000						
DML MAR	1.000	-0.116	0.029	0.120	0.031	0.022
DML IV	1.000	0.006	0.041	0.041	0.045	0.965

Notes: column 'true' shows the true effect, 'bias' the bias of the respective estimator, 'sd' the standard deviation, and 'RMSE' the root mean squared error. Column 'meanSE' displays the average standard error based on the asymptotic approximation across all simulations, 'coverage' the coverage rate of the true effect based on 95% confidence intervals.

In a second simulation design, we set $\gamma = 1$ and $\sigma^2_{U,V} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, such that selection is nonignorable, i.e. related to unobservables as discussed in Section 3, due to a strong correlation of U and V. Table 2 presents the results. DML MAR is no longer unbiased, while the bias of DML IV appears to approach zero as the sample size increases, at the price of somewhat higher standard deviation than DML MAR. However, DML IV dominates DML MAR under either sample size in terms of having a lower RMSE and has thus a more favorable bias-variance trade-off in the scenario considered. While coverage is quite satisfactory for DML IV, the 95% confidence interval mostly fails to include the true effect in the case of DML MAR, in particular

under the larger sample size.

7 Application

As an empirical illustration, we apply our method to the Job Corps (JC) training program. The data come from the National Job Corps Study (NJCS), a randomized social experiment conducted in the mid-to-late 1990s in the United States to evaluate the effectiveness of JC on different labor market outcomes. The JC is the largest and most comprehensive job training program for disadvantaged youth in the US, in which participants are exposed to different types of academic and vocational instruction. The data set contains very detailed pre-treatment information about program participants, such as: expectations, motivations for applying to JC, age, gender, number of children at the moment of treatment assignment, occupation, household income, hourly wage, educational level, marital status, whether the individual was previously attending a school, JC program, or some other academic or vocational training, health status, past employment, types of crimes committed, family support for attending the training, and information on the mother and the father (e.g. education and employment). Furthermore, a large range of variables for instance related to labor market status, employment, income, and education are reassessed in several follow-up interviews after JC assignment.

Schochet, Burghardt, and Glazerman (2001) and Schochet, Burghardt, and McConnell (2008) evaluate the impact of random program assignment on a wide range of labor market outcomes, showing positive effects on education, employment, and earnings in the longer run. Several studies focus on more specific program aspects of JC, e.g. on the effect of the length of exposure to training or of discrete sequences of training interventions on labor market outcomes like employment and earnings, as well as on JC's causal mechanisms, i.e. direct and indirect effects (operating via specific mediating variables) on labor market outcomes and health. Flores, Flores-Lagunes, Gonzales, and Neuman (2012), for instance, acknowledge the existence of several types of instruction, which, along with the self-paced nature of the program, creates selective heterogeneity in the number of weeks the participants are exposed to vocational or academic training. Considering continuously distributed treatment doses while controlling for baseline covariates, they find a positive effect of weeks in training on earnings, however, with decreasing marginal returns as a function of weeks already accomplished. Flores and Flores-Lagunes

(2009) and Huber (2014a) investigate the causal mechanisms underlying JC when considering work experience or employment as mediators, respectively, and find positive direct effects on earnings and general health, respectively, when invoking a selection-on-observables assumption. Flores and Flores-Lagunes (2010) avoid the latter assumption by suggesting a partial identification strategy based on which they estimate bounds the on causal mechanisms of JC when considering the achievement of a GED, high school degree, or vocational degree as mediators. Under their strongest set of bounding assumptions, the results suggest a positive effect on labor market outcomes even net of the indirect mechanism via obtaining a degree. Frölich and Huber (2017) base their analysis of causal mechanisms on an instrumental variable approach and find a positive indirect effect of JC training on earnings through an increase in the number of hours worked.

Estimation approaches specifically dealing with truncated outcomes such as wages, which are only observed and defined for a selective subpopulation like those in employment, have also been considered. For instance, Frumento, Mealli, Pacini, and Rubin (2012) and Zhang, Rubin, and Mealli (2009) consider the principal stratification approach of Frangakis and Rubin (2002) to evaluate the effects of JC on employment and wages among specific groups, e.g. those finding employment irrespective of training participation, rather among the full population. Frumento, Mealli, Pacini, and Rubin (2012) simultaneously address the several identification issues related to noncompliance of training participation with JC assignment as well as missingness in wage outcomes due to survey non-response or non-employment based on a likelihood-based analysis using finite mixture models. They find positive effects on wages and evidence that the program should ideally have been designed differently for different subgroups of individuals depending on their personal characteristics.

Lee (2009) considers a partial identification approach for bounding the effect of JC assignment on wage among those finding employment irrespective of the assignment. Rather than invoking MAR or IV assumptions for sample selection, this method merely relies on the monotonicity of employment in JC assignment (such that being randomized in never decreases the employment state), at the cost of giving up point identification. Semenova (2020) suggests a DML approach to tighten the bounds by controlling for covariates X in a data-driven way. The results based on bounding generally point to a positive (intention-to-treat) effect of JC assignment on wages. Finally, Bodory, Huber, and Laffé rs (2020) consider dynamic treatment evaluation based on

DML, imposing a sequential conditional independence assumption to analyze discrete sequences of training in the first and second year after JC assignment. Controlling for both baseline characteristics and covariates measured after the first treatment period (i.e. one year after JC assignment) they find a positive effect of a sequence of vocational training on employment.

For our empirical analysis we similarly to Frölich and Huber (2017) consider female applicants to JC and aim at estimating the effects of academic or vocational training received in the first year of the program (D) on hourly wage (Y) in the short run measured in the last week of the first year or in the longer run, measured 4 years after random assignment to JC. Hourly wage is only observed conditional on employment (S) in the respective outcome period. Even though JC assignment is random, actual participation in training activities is likely selective and associated with individual factors, similarly to the selection into employment. As for example discussed in Lechner and Wunsch (2013) and Biewen, Fitzenberger, Osikominu, and Paul (2014), the previous labor market history and socio-economic characteristics are likely important confounders when assessing the impact of training interventions, which motivates our DML approach to account for a rich set of covariates in a data-driven way. To assess the short-run effects, we either assume MAR as discussed in Section 2 and use our DML approach based on Theorem 1 to control for our all in all 355 baseline covariates (X), or we impose the IV assumptions of Section 3 to estimate the ATE based on Theorem 3, considering the number of young children in the household at JC assignment as instrument (Z) for employment. Even though numerous studies in labor economics consider children as instrument for employment, the presence of small children might arguably be associated with personal characteristics also affecting the wage and thus violate IV validity - a concern we aim to mitigate by including a rich set of individual pre-treatment characteristics in X that is likely associated with both fertility and wages.

For assessing the longer run effects, we invoke the sequential conditional independence assumption of Section 4 to apply DML based on Theorem 4. To this end, we additionally control for 619 and 156 post-treatment covariates (M) in the second and third year, respectively, after JC assignment, which include detailed information on the labor market participation after the first and prior to the second treatment. Appendix C presents descriptive statistics for selected variables in X and M. We also refer to Bodory, Huber, and Laffers (2020) for a more detailed description of the pre-and post-treatment covariates used in our application and note that all numeric variables have been standardized to have a mean equal to 0 and standard deviation

equal to 0.5 to facilitate the machine learning-based estimation of the nuisance parameters. For estimation, we apply the *treatselDML* and *dyntreatDML* commands of the *causalweight* package for R, using 3-fold cross-fitting and the random forest (with default options of the *SuperLearner* package) as machine learner. The latter is a nonparametric approach allowing for nonlinear associations between the outcome, treatment, and selection on the one hand and the covariates on the other hand.

Table 3: Treatment distribution

treatment	observations
randomized out of JC	1698
controls (no training)	200
academic training	830
vocational training	843

Table 3 reports the total number of females randomized into JC for whom participation in either vocational (843) or academic training (830) in the first year after program assignment is registered in the data. 200 females did not participate in any JC training activity in the first year and serve as the control group in our analysis. Furthermore, 1698 were randomized out of JC.

Table 4: ATE estimates

$D=1$	$D=0$	ATE	standard error	p-value
Theorem 1 (MAR)				
academic	no training	-0.683	1.073	0.524
vocational	no training	0.611	0.629	0.331
Theorem 3 (IV)				
academic	no training	-0.631	1.052	0.549
vocational	no training	0.586	0.645	0.364
Theorem 4 (sequential)				
academic	no training	0.149	0.199	0.454
vocational	no training	0.567	0.208	0.007

Table 4 reports the ATE estimates for academic and vocational training based on our various DML approaches. The upper panel provides the short-run effects on hourly wages in the last week of the first year when assuming MAR. The point estimate of academic training is negative (-0.683 US \$), while that of vocational training is positive (0.611 US \$), but neither effect is statistically significant at any conventional level. The findings are very similar when considering the IV-based estimates shown in the intermediate panel, with negative and positive effects for

the academic and vocational training, respectively, which are again not statistically significant. The lower panel provides the longer-run effects on hourly wages 4 years after assignment based on the sequential conditional independence assumption. While now both ATE estimates are positive, only the effect of vocational training, which amounts to an hourly increase of 0.567 $, is highly statistically significant. Our findings therefore suggest that JC-based education may facilitate human capital accumulation in a way that increases hourly wages after several years, in particular through vocational training, while there is no clear-cut evidence for short term effects.

8 Conclusion

In this paper, we discussed the evaluation of average treatment effects in the presence of sample selection or outcome attrition based on double machine learning. In terms of identifying assumptions, we imposed a selection-on-observables assumption on treatment assignment, which was combined with either selection-on-observables or instrumental variable assumptions concerning the outcome attrition/sample selection process. We also considered a sequential selection-on-observables assumption allowing for dynamic confounding such that covariates jointly affecting the outcome and sample selection may be affected by the treatment, which avoids exclusively relying on pre-treatment covariates. We proposed doubly robust score functions and formally showed the satisfaction of Neyman orthogonality, implying that estimators based on these score functions are robust to moderate (local) regularization biases in the machine learning-based estimation of the outcome, treatment, or sample selection models. Furthermore, we demonstrated the root-n consistency and asymptotic normality of our double machine learning approach to average treatment effect estimation under specific regularity conditions. We also provided an empirical illustration to the US Job Corps data, in which we assessed the effects of training on hourly wage one and four years after program assignment and found some statistical evidence for positive longer-run impacts. Our estimation procedure is available in the *causalweight* package for the statistical software R.

References

Abowd, J., B. Crepon, and F. Kramarz (2001): "Moment Estimation With Attrition: An Application to Economic Models," *Journal of the American Statistical Association*, 96, 1223–1230.

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Angrist, J., E. Bettinger, and M. Kremer (2006): "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *American Economic Review*, 96, 847–862.

Bang, H., and J. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

Barnwell, J.-L., and S. Chaudhuri (2020): "Efficient estimation in sub and full populations with monotonically missing at random data," *working paper, McGill University, Montreal.*

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Biewen, M., B. Fitzenberger, A. Osikominu, and M. Paul (2014): "The Effectiveness of Public-Sponsored Training Revisited: The Importance of Data and Methodological Choices," *Journal of Labor Economics*, 32), pages = 837-897,.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, 71, 655–679.

Bodory, H., and M. Huber (2018): "The causalweight package for causal inference in R," *SES Working Paper 493, University of Fribourg.*

Bodory, H., M. Huber, and L. Lafférs (2020): "Evaluating (weighted) dynamic treatment effects by double machine learning," *arXiv preprint arXiv:2012.00370.*

Carroll, R., D. Ruppert, and L. Stefanski (1995): *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

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

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

d'Haultfoeuille, X. (2010): "A new instrumental method for dealing with endogenous selection," *Journal of Econometrics*, 154, 1–15.

Farrell, M. H., T. Liang, and S. Misra (2018): "Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands," *working paper, University of Chicago.*

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998): "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources*, 33, 251–299.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237.*

Flores, C. A., and A. Flores-Lagunes (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida.*

Flores, C. A., A. Flores-Lagunes, A. Gonzales, and T. Neuman (2012): "Estimating the effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps," *The Review of Economics and Statistics*, 94, 153–171.

Frangakis, C., and D. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Frölich, M., and M. Huber (2017): "Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society: Series(B)*, 79, 1645–1666.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012): "Evaluating the Effect of

Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data," *Journal of the American Statistical Association*, 107, 450–466.

Gronau, R. (1974): "Wage comparisons-a selectivity bias," *Journal of Political Economy*, 82, 1119–1143.

Hausman, J., and D. Wise (1979): "Attrition Bias In Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47(2), 455–473.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Huber, M. (2012): "Identification of average treatment effects in social experiments under alternative forms of attrition," *Journal of Educational and Behavioral Statistics*, 37, 443–474.

Huber, M. (2014a): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

Huber, M. (2014b): "Treatment evaluation in the presence of sample selection," *Econometric Reviews*, 33, 869–905.

Huber, M., and B. Melly (2015): "A Test of the Conditional Independence Assumption in Sample Selection Models," *Journal of Applied Econometrics*, 30, 1144–1168.

Imai, K. (2009): "Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment," *Journal of the Royal Statistical Society Series C*, 58, 83–104.

Imbens, G. W. (2004): "Nonparametric estimation of average treatment effects under exogeneity: a review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Lechner, M. (2009): "Sequential Causal Models for the Evaluation of Labor Market Programs," *Journal of Business and Economic Statistics*, 27, 71–83.

Lechner, M., and C. Wunsch (2013): "Sensitivity of matching-based program evaluations to the availability of control variables," *Labour Economics*, 21, 111–121.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Levy, J. (2019): "Tutorial: Deriving The Efficient Influence Curve for Large Models," *arXiv preprint arXiv:1903.01706*.

Little, R., and D. Rubin (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

Little, R. J. A. (1995): "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Luo, Y., and M. Spindler (2016): "High-Dimensional L_2Boosting: Rate of Convergence," .

Negi, A. (2020): "Doubly weighted M-estimation for nonrandom assignment and missing outcomes," *arXiv preprint arXiv:2011.11485*.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Newey, W. K. (2007): "Nonparametric continuous/discrete choice models," *International Economic Review*, 48, 1429–1439.

Neyman, J. (1959): *Optimal asymptotic tests of composite statistical hypotheses*p. 416–444. Wiley.

Robins, J. (1986): "A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J., A. Rotnitzky, and L. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

Robins, J. M. (1998): "Marginal Structural Models," in *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, p. 1–10.

Robins, J. M., S. D. Mark, and W. K. Newey (1992): "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479–495.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

Rubin, D. (1980): "Comment on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test' by D. Basu," *Journal of American Statistical Association*, 75, 591–593.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Schochet, P., J. Burghardt, and S. Glazerman (2001): "National Job Corps Study: The Impacts of Job Corps on Participants Employment and Related Outcomes," *Report, Washington, DC: Mathematica Policy Research, Inc.*

Schochet, P., J. Burghardt, and S. McConnell (2008): "Does Job Corps Work? Impact Findings from the National Job Corps Study," *The American Economic Review*, 98, 1864–1886.

Semenova, V. (2020): "Better Lee Bounds," *arXiv preprint arXiv:2008.12720*.

Shah, A., N. Laird, and D. Schoenfeld (1997): "A Random-Effects Model for Multiple Characteristics With Possibly Missing Data," *Journal of the American Statistical Association*, 92, 775–779.

Sloczyński, T., and J. M. Wooldridge (2018): "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 34, 112–133.

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

Wager, S., and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wooldridge, J. (2002): "Inverse Probability Weigthed M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 141–162.

——— (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

Zhang, J., D. Rubin, and F. Mealli (2009): "Likelihood-based analysis of causal effects of job-training programs using principal stratification," *Journal of the American Statistical Association*, 104, 166–176.

Appendices

A Proofs of theorems

For the proofs of Theorems 1, 2, and 3 it is sufficient to verify the conditions of Assumptions 3.1 and 3.2 of Theorems 3.1 and 3.2 as well as Corollary 3.2 in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). All bounds hold uniformly over $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws, and we omit P for brevity.

A.1 Proof of Theorem 1

Define the nuisance parameters to be the vector of functions $\eta = (p_d(X), \pi(D,X), \mu(D,S,X))$, with $p_d(X) = \Pr(D = d|X)$, $\pi(D,X) = \Pr(S = 1|D,X)$, and $\mu(D,S,X) = E[Y|D,S,X]$. The Neyman-orthogonal score function for the counterfactual $\Psi_{d0} = E[Y(d)]$ is given by the following expression, with $W = (Y \cdot S, D, S, X)$:

$$\psi_d(W, \eta, \Psi_{d0}) \quad = \quad \frac{I\{D = d\} \cdot S \cdot [Y - \mu(d,1,X)]}{p_d(X) \cdot \pi(d,X)} + \mu(d,1,X) - \Psi_{d0}. \tag{A.1}$$

Let $\mathcal{T}_n$ be the set fo all $\eta = (p_d, \pi, \mu)$ consisting of P-square integrable functions p_d, π, μ such that

$$\begin{aligned}
\|\eta - \eta_0\|_q &\leq C, \\
\|\eta - \eta_0\|_2 &\leq \delta_n, \\
\|p_d(X) - 1/2\|_\infty &\leq 1/2 - \epsilon, \\
\|\pi(D,X) - 1/2)\|_\infty &\leq 1/2 - \epsilon, \\
\|\mu(D,S,X) - \mu_0(D,S,X)\|_2 \times \|p_d(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}, \\
\|\mu(D,S,X) - \mu_0(D,S,X)\|_2 \times \|\pi(D,X) - \pi_0(D,X)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned} \tag{A.2}$$

We furthermore replace the sequence $(\delta_n)_{n\geq 1}$ by $(\delta'_n)_{n\geq 1}$, where $\delta'_n = C_\epsilon \max(\delta_n, n^{-1/2})$, where C_ϵ is sufficiently large constant that only depends on C and ϵ.

Assumption 3.1: Linear scores and Neyman orthogonality

Assumption 3.1(a)

Moment Condition: The moment condition $E\Big[\psi_d(W,\eta_0,\Psi_{d0})\Big] = 0$ holds:

$$\begin{aligned} E\Big[\psi_d(W,\eta_0,\Psi_{d0})\Big] &= E\Bigg[\overbrace{E\left[\frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X)]}{p_{d0}(X)\cdot\pi_0(d,X)}\middle| X\right]}^{=E[[Y-\mu_0(d,1,X)|D=d,S=1,X]=0} + \mu_0(d,1,X) - \Psi_{d0}\Bigg] \\ &= E[\mu_0(d,1,X)] - \Psi_{d0} = 0, \end{aligned}$$

where the first equality follows from the law of iterated expectations.

Assumption 3.1(b) Linearity: The score $\psi_d(W,\eta_0,\Psi_{d0})$ is linear in Ψ_{d0} : $\psi_d(W,\eta_0,\Psi_{d0}) = \psi_d^a(W,\eta_0)\cdot\Psi_0^d + \psi_d^b(W,\eta_0)$ with $\psi_d^a(W,\eta_0) = -1$ and

$$\psi_d^b(W,\eta_0) = \frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X)]}{p_{d0}(X)\cdot\pi_0(d,X)} + \mu_0(d,1,X).$$

Assumption 3.1(c)

Continuity: The expression for the second Gateaux derivative of a map $\eta \mapsto E[\psi_d(W,\eta,\Psi_{d0})]$, given in (A.11), is continuous.

Assumption 3.1(d)

Neyman Orthogonality: For any $\eta \in \mathcal{T}_n$, the Gateaux derivative in the direction $\eta - \eta_0 = (\pi(D,X) - \pi_0(D,X), p^d(X) - p_0^d(X), \mu(D,S,X) - \mu_0(D,S,X)$ is given by:

$$\begin{aligned} &\partial E\big[\psi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big] = \\ &- E\left[\frac{I\{D=d\}\cdot S\cdot[\mu(d,1,X)-\mu_0(d,1,X)]}{p_{d0}(X)\cdot\pi_0(d,X)}\right] \qquad (*) \\ &+ E[\mu(d,1,X)-\mu_0(d,1,X)] \qquad (**) \\ &- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X)]}^{E[\cdot|X]=E[Y-\mu_0(d,1,X)|D=d,S=1,X]=0}}{p_{d0}(X)\cdot\pi_0(d,X)}\cdot\frac{p_d(X)-p_{d0}(X)}{p_{d0}(X)}\right] \\ &- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X)]}^{E[\cdot|X]=0}}{p_{d0}(X)\cdot\pi_0(d,X)}\cdot\frac{\pi(d,X)-\pi_0(d,X)}{\pi_0(d,X)}\right] = 0. \end{aligned}$$

The Gateaux derivative is zero because expressions $(*)$ and $(**)$ cancel out. To see this, note that by the law of

iterated expectations, $(*)$ corresponds to

$$
\begin{aligned}
& -E\left[E\left[\frac{I\{D=d\}}{p_{d0}(X)}\cdot E\left[\frac{S\cdot[\mu(d,1,X)-\mu_0(d,1,X)]}{\pi_0(d,X)}\middle| D=d,X\right]\middle| X\right]\right] \\
= \quad & -E\left[E\left[\frac{I\{D=d\}}{p_{d0}(X)}\cdot \frac{\overbrace{E[S|D=d,X]}^{=\pi_0(d,X)}\cdot[\mu(d,1,X)-\mu_0(d,1,X)]}{\pi_0(d,X)}\middle| X\right]\right] \\
= \quad & -E\left[\frac{\overbrace{E[I\{D=d\}|X]}^{=p_{d0}(X)}}{p^{d0}(X)}\cdot[\mu(d,1,X)-\mu_0(d,1,X)]\right] = -E[\mu(d,1,X)-\mu_0(d,1,X)],
\end{aligned}
$$

Therefore,

$$\partial E\big[\psi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big]=0$$

proving that the score function is orthogonal.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

Assumption 3.2(a)

This assumption directly follows from the construction of the set $\mathcal{T}_n$ and the regularity conditions (Assumption 10).

Assumption 3.2(b)

Bound for m_N:

Consider the following inequality

$$
\begin{aligned}
\|\mu_0(D,S,X)\|_q &= \left(E\left[|\mu_0(D,S,X)|^q\right]\right)^{\frac{1}{q}} \\
&= \left(\sum_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X)|^q\Pr(D=d,S=s|X)\right]\right)^{\frac{1}{q}} \\
&\geq \epsilon^{2/q}\left(\sum_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X)|^q\right]\right)^{\frac{1}{q}} \\
&\geq \epsilon^{2/q}\left(\max_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X)|^q\right]\right)^{\frac{1}{q}} \\
&= \epsilon^{2/q}\left(\max_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} \|\mu_0(d,s,X)\|_q\right),
\end{aligned}
$$

where the first equality follows from definition, the second from the law of total probability, first inequality from the fact that $\Pr(D=d,S=1|X)=p_{d0}(X)\cdot\pi_0(d,X)\geq\epsilon^2$ and $\Pr(D=d,S=0|X)=p_{d0}(X)\cdot(1-\pi_0(d,X))\geq\epsilon^2$. Furthermore, by Jensen's inequality $\|\mu_0(D,S,X)\|_q\leq\|Y\|_q$ and hence $\|\mu_0(d,1,X)\|_q\leq C/\epsilon^{2/q}$ by conditions (A.8). Using similar steps, for any $\eta\in\mathcal{T}_N:\|\mu(d,1,X)-\mu_0(d,1,X)\|_q\leq C/\epsilon^{2/q}$ because $\|\mu(D,S,X)-\mu_0(D,S,X)\|_q\leq C$.

Consider

$$E\Big[\psi_d(W,\eta,\Psi_{d0})\Big] \quad = \quad E\Bigg[\underbrace{\frac{I\{D=d\}\cdot S}{p_d(X)\cdot\pi(d,X)}\cdot Y}_{=I_1}+\underbrace{\Big(1-\frac{I\{D=d\}\cdot S}{p_d(X)\cdot\pi(d,X)}\Big)\mu(d,1,X)}_{=I_2}-\Psi_{d0}\Bigg]$$

and thus

$$\begin{aligned}\|\psi_d(W,\eta,\Psi_{d0})\|_q &\leq \|I_1\|_q+\|I_2\|_q+\|\Psi_{d0}\|_q\\ &\leq \frac{1}{\epsilon^2}\|Y\|_q+\frac{1-\epsilon}{\epsilon}\|\mu(d,1,X)\|_q+|\Psi_{d0}|\\ &\leq C\left(\frac{1}{\epsilon^2}+\frac{2}{\epsilon^{2/q}}\cdot\frac{1-\epsilon}{\epsilon}+\frac{1}{\epsilon}\right),\end{aligned}$$

because of triangular inequality and because the following set of inequalities hold:

$$\begin{aligned}\|\mu(d,1,X)\|_q &\leq \|\mu(d,1,X)-\mu_0(d,1,X)\|_q+\|\mu_0(d,1,X)\|_q\leq 2C/\epsilon^{2/q}, \qquad (A.3)\\ |\Psi_0^{d_2}| &= |E[\mu_0(d,1,X)]|\leq E\Big[\,|\mu_0(d,1,X)|^1\,\Big]^{\frac{1}{1}}=\|\mu_0(d,1,X)\|_{P,1}\\ &\leq \|\mu_0(d,1,X)\|_2\leq\|Y\|_2/\epsilon^{2/2}\overbrace{\leq}^{q>2}\|Y\|_q/\epsilon\leq C/\epsilon.\end{aligned}$$

which gives the upper bound on m_n in Assumption 3.2(b) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

Bound for m_n':

Notice that

$$\Big(E[|\psi_d^a(W,\eta)|^q]\Big)^{1/q}=1$$

and this gives the upper bound on m_N' in Assumption 3.2(b).

Assumption 3.2(c)

Bound for r_n:

For any $\eta=(p_d,\pi,\nu)$ we have

$$\Big|E\Big(\psi_d^a(W,\eta)-\psi_{d0}^a(W,\eta_0)\Big)\Big|=|1-1|=0\leq\delta_N',$$

and thus we have the bound on r_n from Assumption 3.2(c).

In the following, we omit arguments for the sake of brevity and use $p_d=p_d(X),\pi=\pi(d,X),\mu=\mu(d,1,X)$ and similarly for p_{d0},π_0,μ_0.

Bound for r_n':

$$\begin{aligned}
&\|\psi_d(W,\eta,\Psi_{d0}) - \psi_d(W,\eta_0,\Psi_{d0})\|_2 \leq \left\| I\{D=d\}\cdot S\cdot Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2 \qquad \text{(A.4)}\\
+\quad &\left\| I\{D=d\}\cdot S\cdot\left(\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right)\right\|_2 + \|\mu-\mu_0\|_2\\
\leq\quad &\left\| Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2 + \left\|\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right\|_2 + \|\mu-\mu_0\|_2\\
\leq\quad &\frac{C}{\epsilon^4}\delta_n\left(1+\frac{1}{\epsilon}\right)+\delta_n\left(\frac{1}{\epsilon^5}+C+\frac{C}{\epsilon}\right)+\frac{\delta_n}{\epsilon}\leq\delta_n'
\end{aligned}$$

as long as C_ϵ in the definition of δ_n' is sufficiently large. This gives the bound on r_n' from Assumption 3.2(c). Here we made use of the fact that $\|\mu-\mu_0\|_2 = \|\mu(d,1,X)-\mu_0(d,1,X)\|_2 \leq \delta_n/\epsilon$, and $\|\pi-\pi_0\|_2 = \|\pi(d,X)-\pi_0(d,X)\|_2 \leq \delta_n/\epsilon$ using similar steps as in Assumption 3.1(b).

The last inequality in (A.10) holds because for the first term we have

$$\begin{aligned}
&\left\| Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2 \leq C\left\|\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right\|_2 \leq \frac{C}{\epsilon^4}\|p_{d0}\pi_0-p_d\pi\|_2\\
=\quad &\frac{C}{\epsilon^4}\|p_{d0}\pi_0-p_d\pi+p_{d0}\pi-p_{d0}\pi\|_2 \leq \frac{C}{\epsilon^4}\left(\|p_{d0}(\pi_0-\pi)\|_2+\|\pi_0(p_{d0}-p_d)\|_2\right)\\
\leq\quad &\frac{C}{\epsilon^4}\left(\|\pi_0-\pi\|_2+\|p_{d0}-p_d\|_2\right)\leq\frac{C}{\epsilon^4}\delta_n\left(1+\frac{1}{\epsilon}\right),
\end{aligned}$$

where the first inequality follows from the second inequality in Assumption 4(a). The second term in (A.10) is bounded by

$$\begin{aligned}
&\left\|\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right\|_2 \leq \frac{1}{\epsilon^4}\|p_{d0}\pi_0\mu-p_d\pi\mu_0\|_2 = \frac{1}{\epsilon^4}\|p_{d0}\pi_0\mu-p_d\pi\mu_0+p_{d0}\pi_0\mu_0-p_{d0}\pi_0\mu_0\|_2\\
\leq\quad &\frac{1}{\epsilon^4}\left(\|p_{d0}\pi_0(\mu-\mu_0)\|_2+\|\mu_0(p_{d0}\pi_0-p_d\pi)\|_2\right)\leq\frac{1}{\epsilon^4}\left(\|\mu-\mu_0\|_2+C\|p_{d0}\pi_0-p_d\pi\|_2\right)\\
\leq\quad &\frac{1}{\epsilon^4}\left(\frac{\delta_n}{\epsilon}+C\|p_{d0}\pi_0-p_d\pi\|_2\right)\leq\delta_n\left(\frac{1+C}{\epsilon^5}+\frac{C}{\epsilon^4}\right),
\end{aligned}$$

where the third inequality follows from $E[Y^2|D=d,S=1,X] \geq (E[Y|D=d,S=1,X])^2 = \mu_0^2(d,1,X)$ by conditional Jensen's inequality and therefore $\|\mu_0(d,1,X)\|_\infty \leq C^2$.

Bound for λ_n':

Now consider

$$f(r) := E[\psi_d(W;\Psi_{d0},\eta+r(\eta-\eta_0)]$$

For any $r\in(0,1)$:

$$
\begin{aligned}
\frac{\partial^2 f(r)}{\partial r^2} &= E\left[2\cdot I\{D=d\}\cdot S\cdot(Y-\mu_0-r(\mu-\mu_0))\frac{(p_d-p_{d0})^2}{(p_{d0}+r(p_d-p_{d0}))^3(\pi_0+r(\pi-\pi_0))}\right] \\
&+ E\left[2\cdot I\{D=d\}\cdot S\cdot(Y-\mu_0-r(\mu-\mu_0))\frac{(\pi-\pi_0)^2}{(p_{d0}+r(p_d-p_{d0}))(\pi_0+r(\pi-\pi_0))^3}\right] \\
&+ E\left[2\cdot I\{D=d\}\cdot S\cdot(Y-\mu_0-r(\mu-\mu_0))\frac{(p_d-p_{d0})(\pi-\pi_0)}{(p_{d0}+r(p_d-p_{d0}))^2(\pi_0+r(\pi-\pi_0))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\cdot S\cdot(\mu-\mu_0)\frac{(p_d-p_{d0})(\pi_0+r(\pi-\pi_0))}{(p_{d0}+r(p_d-p_{d0}))^2(\pi_0+r(\pi-\pi_0))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\cdot S\cdot(\mu-\mu_0)\frac{(p_{d0}+r(p_d-p_{d0}))(\pi-\pi_0)}{(p_{d0}+r(p_d-p_{d0}))^2(\pi_0+r(\pi-\pi_0))^2}\right]
\end{aligned}
\tag{A.5}
$$

Note that because

$$
\begin{aligned}
E[Y-\mu_0(d,1,X)|D=d,S=1,X] &= 0, \\
|p_d-p_{d0}|\leq 2, \qquad |\pi-\pi_0| &\leq 2 \\
\|\mu_0\|_q \leq \|Y\|_q/\epsilon^{1/q} &\leq C/\epsilon^{2/q} \\
\|\mu-\mu_0\|_2\times\|p_d-p_{d0}\|_2 &\leq \delta_n n^{-1/2}/\epsilon, \\
\|\mu-\mu_0\|_2\times\|\pi-\pi_0\|_2 &\leq \delta_n n^{-1/2}/\epsilon^2,
\end{aligned}
$$

we get that for some constant C''_ϵ that only depends on C and ϵ

$$
\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \leq C''_\epsilon \delta_n n^{-1/2} \leq \delta'_n n^{-1/2}
$$

and this gives the upper bound on λ'_n in Assumption 3.2(c) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) as long as $C_\epsilon \geq C''_\epsilon$. We used the following inequalities

$$
\begin{aligned}
\|\mu-\mu_0\|_2 &= \|\mu(d,1,X)-\mu_0(d,1,X)\|_2 \leq \|\mu(D,S,X)-\mu_0(D,S,X)\|_2/\epsilon \\
\|\pi-\pi_0\|_2 &= \|\pi(d,X)-\pi_0(d,X)\|_2 \leq \|\pi(D,X)-\pi_0(D,X)\|_2/\epsilon,
\end{aligned}
$$

and these can be shown using similar steps as in Assumption 3.1(b).

To verify that $\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \leq C''_\epsilon \delta_n n^{-1/2}$ holds, note that by the triangular inequality it is sufficient to bound the absolute value of each of the ten terms in (A.11) separately. We illustrate it for the first and last terms. For the

first term:

$$\begin{aligned}
&\left|E\left[2\cdot I\{D=d\}\cdot S\cdot(Y-\mu_0-r(\mu-\mu_0))\frac{(p_d-p_{d0})^2}{(p_{d0}+r(p_d-p_{d0}))^3(\pi_0+r(\pi-\pi_0))}\right]\right| \\
\leq\ & \frac{2}{\epsilon^4}\left|E\left[I\{D=d\}\cdot S\cdot(Y-\mu_0-r(\mu-\mu_0))(p_d-p_{d0})^2\right]\right| \\
\leq\ & \frac{8}{\epsilon^4}\left|E\left[I\{D=d\}\cdot S\cdot(Y-\mu_0)\right]\right|+\frac{2}{\epsilon^4}\left|E\left[r(\mu-\mu_0)(p_d-p_{d0})^2\right]\right| \\
\leq\ & \frac{2\cdot 2}{\epsilon^4}\left|E\left[1\cdot(\mu-\mu_0)(p_d-p_{d0})\right]\right|\leq\frac{4}{\epsilon^4}\frac{\delta_n}{\epsilon}n^{-1/2}.
\end{aligned}$$

For the second inequality we used the fact that for $1\geq p_{d0}+r(p_d-p_{d0})=(1-r)p_{d0}+rp_d\geq(1-r)\epsilon+r\epsilon=\epsilon$ and similarly for π and in the third Holder's inequality. Bounding of the second and third terms follows similarly.

For the fourth term, we get

$$\begin{aligned}
&\left|E\left[2\cdot I\{D=d\}\cdot S\cdot(\mu-\mu_0)\frac{(p_d-p_{d0})(\pi_0+r(\pi-\pi_0))}{(p_{d0}+r(p_d-p_{d0}))^2(\pi_0+r(\pi-\pi_0))^2}\right]\right| \\
\leq\ & \frac{2}{\epsilon^4}\left|E\left[I\{D=d\}\cdot S\cdot(\mu-\mu_0)(p_d-p_{d0})\right]\right|\leq\frac{2}{\epsilon^4}\frac{\delta_n}{\epsilon}n^{-1/2}
\end{aligned}$$

where in addition we made use of conditions (A.8). The last term is bounded similarly.

Assumption 3.2(d)

$$\begin{aligned}
E\Big[(\psi^d(W,\eta_0,\Psi_{d0}))^2\Big] &= E\left[\left(\underbrace{\frac{I\{D=d\}\cdot S\cdot[Y-\mu_0]}{p_{d0}\cdot\pi_0}}_{=I_1}+\underbrace{\mu_0-\Psi_{d0}}_{=I_2}\right)^2\right] \\
&= E[I_1^2+I_2^2]\geq E[I_1^2] \\
&= E\left[I\{D=d\}\cdot S\cdot\left(\frac{[Y-\mu_0]}{p_{d0}\cdot\pi_0}\right)^2\right] \\
&\geq \epsilon^2E\left[\left(\frac{[Y-\mu_0]}{p_{d0}\cdot\pi_0}\right)^2\right] \\
&\geq \frac{\epsilon^2c^2}{(1-\epsilon)^4}>0,
\end{aligned}$$

because $\Pr(D=d,S=1|X)=p_{d0}(X)\cdot\pi_0(d,X)\geq\epsilon^2$, $p_{d0}(X)\leq 1-\epsilon$ and $\pi_0(d,X)\leq 1-\epsilon$.

The second equality follows from

$$E\Big[I_1\cdot I_2\Big] = E\left[\overbrace{\frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X)]}{p_{d0}(X)\cdot\pi_0(d,X)}}^{E[\cdot|X]=0}\cdot[\mu_0(d,1,X)-\Psi_{d0}]\right].$$

A.2 Proof of Theorem 2

Define the nuisance parameters to be the vector of functions $\eta=(\pi(D,X,Z),p_d(X,\Pi),\mu(D,S,X,\Pi))$, with $\Pi=\pi(D,X,Z)=\Pr(S=1|D,X,Z)$, $p_d(X,\Pi)=\Pr(D=d|X,\pi(D,X,Z))$, and $\mu(D,S,X,\Pi)=E[Y|D,S,X,\pi(D,X,Z)]$.

The shrinking neighbourhood $\mathcal{T}_n^*$ of nuisance parameter vector $\eta=(\pi,p_d,\mu)$ is defined analogously to $\mathcal{T}_n$ from (A.8) in the proof of theorem 1.

The score function for the counterfactual $\Psi_{d0}^{S=1} = E[Y(d)|S=1]$ is given by:

$$\phi_{d,S=1}(W, \eta, \Psi_{d0}^{S=1}) \quad = \quad \frac{I\{D=d\} \cdot [Y - \mu(d,1,X,\Pi)]}{p_d(X)} + \mu(d,1,X,\Pi) - \Psi_{d0}^{S=1}. \tag{A.6}$$

Assumption 3.1: Linear scores and Neyman orthogonality

Assumption 3.1(a)

Moment Condition: The moment condition $E\Big[\phi_{d,S=1}(W, \eta_0, \Psi_{d0}^{S=1})|S=1\Big] = 0$ holds:

$$\begin{aligned} E\Big[\phi_{d,S=1}(W, \eta_0, \Psi_{d0}^{S=1})\Big|S=1\Big] \quad &= \quad E\Bigg[\overbrace{E\left[\frac{I\{D=d\} \cdot [Y - \mu_0(d,1,X,\Pi_0)]}{p_{d0}(X,\Pi_0)}\middle| S=1, X, \Pi_0\right]}^{=E[Y-\mu_0(d,1,X,\Pi_0)|D=d,S=1,X,\Pi_0]=0} \\ &+ \quad \mu_0(d,1,X,\Pi_0) - \Psi_{d0}^{S=1}\Bigg|S=1\Bigg] \\ &= \quad E[\mu_0(d,1,X,\Pi_0)|S=1] - \Psi_{d0}^{S=1} = 0, \end{aligned}$$

where the first equality follows from the law of iterated expectations.

Assumption 3.1(b) Linearity: The score $\phi_{d,S=1}(W, \eta_0, \Psi_{d0}^{S=1})$ is linear in $\Psi_{d0}^{S=1}$: $\phi_{d,S=1}(W, \eta_0, \Psi_{d0}^{S=1}) = \phi_{d,S=1}^a(W, \eta_0) \cdot \Psi_{d0}^{S=1} + \phi_{d,S=1}^b(W, \eta_0)$ with $\phi_{d,S=1}^a(W, \eta_0) = -1$ and

$$\phi_{d,S=1}^b(W, \eta_0) \quad = \quad \frac{I\{D=d\} \cdot [Y - \mu_0(d,1,X,\Pi_0)]}{p_{d0}(X,\Pi_0)} + \mu_0(d,1,X,\Pi_0).$$

Assumption 3.1(c)

Continuity: The expression for the second Gateaux derivative of a map $\eta \mapsto E[\phi_{d,S=1}(W, \eta, \Psi_{d0}^{S=1})]$, given in (A.6), is continuous.

Assumption 3.1(d)

Neyman Orthogonality: For any $\eta \in \mathcal{T}_n$, the Gateaux derivative in the direction $\eta - \eta_0 = (\pi(D,X,Z) -$

$\pi_0(D,X,Z), p_d(X,\Pi) - p_{d0}(X,\Pi), \mu(D,S,X,\Pi) - \mu_0(D,S,X,\Pi))$ is given by:

$$
\begin{aligned}
&\partial E\big[\phi_{d,S=1}(W,\eta,\Psi_d^{S=1})|S=1\big]\big[\eta-\eta_0\big] = \\
&- E\left[\frac{I\{D=d\}\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]}{p_{d0}(X,\pi_0(d,X,Z))}\middle| S=1\right] && (*)\\
&+ E[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))|S=1] && (**)\\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot[Y-\mu_0(d,1,X,\pi_0(d,X,Z))]}^{E[\cdot|S=1,X,\Pi_0]=E[Y-\mu_0(d,1,X,\Pi_0)|D=d,S=1,X,\Pi_0]=0}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot\frac{p_d(X,\pi_0(d,X,Z)) - p_{d0}(X,\pi_0(d,X,Z))}{p_{d0}(X,\pi_0(d,X,Z))}\middle| S=1\right]\\
&- E\left[\frac{I\{D=d\}\cdot\partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]}{p_{d0}(X,\pi_0(d,X,Z))}\middle| S=1\right] && (***)\\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot[Y-\mu_0(d,1,X,\pi_0(d,X,Z))]}^{E[\cdot|S=1,X,\Pi_0]=E[Y-\mu_0(d,1,X,\Pi_0)|D=d,S=1,X,\Pi_0]=0}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot\frac{\partial E[p_{d0}^2(X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]}{p_{d0}(X,\pi_0(d,X,Z))}\middle| S=1\right]\\
&+ \partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)|S=1] && (****)\\
&= 0.
\end{aligned}
$$

The Gateaux derivative is zero because expressions $(*)$ and $(**)$ as well as $(***)$ and $(****)$, respectively, cancel out. To see this, note that by the law of iterated expectations and the fact that conditioning on D, X, Π is equivalent conditioning on D, X, Π (because Π is deterministic in Z conditional on D, X), $(*)$ corresponds to

$$
\begin{aligned}
&-E\left[\frac{\overbrace{E[I\{D=d\}|X,\Pi_0]}^{=p_{d0}(X,\pi_0(d,X,Z))}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]\middle| S=1\right]\\
= \quad &-E[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))|S=1],
\end{aligned}
$$

which cancels out with $(**)$. In an analogous way, it can be shown that $(***)$ corresponds to

$$E[-\partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]|S=1],$$

which cancels out with $(****)$. Therefore,

$$\partial E\big[\phi_{d,S=1}(W,\eta,\Psi_d^{S=1})\big]\big[\eta-\eta_0\big] = 0$$

proving that the score function is orthogonal.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

This proof follows in a similar way as the proof of Theorem 1 and is omitted for brevity.

A.3 Proof of Theorem 3

The score function for the counterfactual $\Psi_{d0} = E[Y(d)]$ is given by:

$$\phi_d(W, \eta, \Psi_{d0}) \quad = \quad \frac{I\{D=d\} \cdot S \cdot [Y - \mu(d, 1, X, \Pi)]}{p_d(X) \cdot \pi(d, X, Z)} + \mu(d, 1, X, \Pi) - \Psi_{d0}. \tag{A.7}$$

Assumption 3.1: Linear scores and Neyman orthogonality

Assumption 3.1(a)

Moment Condition: The moment condition $E\Big[\psi_d(W, \eta_0, \Psi_{d0})\Big] = 0$ holds:

$$\begin{aligned} E\Big[\phi_d(W, \eta_0, \Psi_{d0})\Big] &= E\Bigg[\overbrace{E\left[\frac{I\{D=d\} \cdot S \cdot [Y - \mu_0(d, 1, X, \Pi_0)]}{p_{d0}(X, \Pi_0) \cdot \pi_0(d, X, Z)} \middle| X, \Pi_0\right]}^{=E[Y-\mu_0(d,1,X,\Pi_0)|D=d,S=1,X,\Pi_0]=0} + \mu_0(d, 1, X, \Pi_0) - \Psi_{d0}\Bigg] \\ &= E[\mu_0(d, 1, X, \Pi_0)] - \Psi_{d0} = 0, \end{aligned}$$

where the first equality follows from the law of iterated expectations.

Assumption 3.1(b) Linearity: The score $\phi_d(W, \eta_0, \Psi_{d0})$ is linear in Ψ_{d0} : $\phi_d(W, \eta_0, \Psi_{d0}) = \phi_d^a(W, \eta_0) \cdot \Psi_{d0} + \phi_d^b(W, \eta_0)$ with $\phi_d^a(W, \eta_0) = -1$ and

$$\phi_d^b(W, \eta_0) \quad = \quad \frac{I\{D=d\} \cdot S \cdot [Y - \mu_0(d, 1, X, \Pi_0)]}{p_{d0}(X, \Pi_0) \cdot \pi_0(d, X, Z)} + \mu_0(d, 1, X, \Pi_0).$$

Assumption 3.1(c)

Continuity: The expression for the second Gateaux derivative of a map $\eta \mapsto E[\phi_d(W, \eta, \Psi_{d0})]$, given in (A.7), is continuous.

Assumption 3.1(d)

Neyman Orthogonality: For any $\eta \in \mathcal{T}_n$, the Gateaux derivative in the direction $\eta - \eta_0 = (\pi(D,X,Z) - \pi_0(D,X,Z), p_d(X,\Pi) - p_{d0}(X,\Pi), \mu(D,S,X,\Pi) - \mu_0(D,S,X,\Pi))$ is given by:

$$
\begin{aligned}
&\partial E\big[\phi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big] = \\
&- E\left[\frac{I\{D=d\}\cdot S\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]}{p_{d0}(X,\pi_0(d,X,Z))\cdot\pi_0(d,X,Z)}\right] && (*)\\
&+ E[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))] && (**)\\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,\pi_0(d,X,Z))]}^{E[\cdot|X,\Pi_0]=E[Y-\mu_0(d,1,X,\pi_0(d,X,Z))|D=d,S=1,X,\Pi_0]=0}}{p_{d0}(X,\pi_0(d,X,Z))\cdot\pi_0(d,X,Z)}\cdot\frac{p_d(X,\pi_0(d,X,Z)) - p_{d0}(X,\pi_0(d,X,Z))}{p_{d0}(X,\pi_0(d,X,Z))}\right]\\
&- E\left[\frac{I\{D=d\}\cdot S\cdot\partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]}{p_{d0}(X,\pi_0(d,X,Z))\cdot\pi_0(d,X,Z)}\right] && (***)\\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,\pi_0(d,X,Z))]}^{E[\cdot|X,\Pi_0]=E[Y-\mu_0(d,1,X,\pi_0(d,X,Z))|D=d,S=1,X,\Pi_0]=0}}{p_{d0}(X,\pi_0(d,X,Z))\cdot\pi_0(d,X,Z)}\cdot\frac{\pi(d,X,Z)-\pi_0(d,X,Z)}{\pi_0(d,X,Z)}\right]\\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,\pi_0(d,X,Z))]}^{E[\cdot|X,\Pi_0]=E[Y-\mu_0(d,1,X,\pi_0(d,X,Z))|D=d,S=1,X,\Pi_0]=0}}{p_{d0}(X,\pi_0(d,X,Z))\cdot\pi_0(d,X,Z)}\cdot\frac{\partial E[p_{d0}^2(X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]}{p_{d0}(X,\pi_0(d,X,Z))}\right]\\
&+ \partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)] && (****)\\
&= 0.
\end{aligned}
$$

The Gateaux derivative is zero because expressions $(*)$ and $(**)$ as well as $(***)$ and $(****)$, respectively, cancel out. To see this, note that by the law of iterated expectations and the fact that conditioning on D,X,Z is equivalent conditioning on D,X,Π (because Π is deterministic in Z conditional on D,X), $(*)$ corresponds to

$$
\begin{aligned}
& -E\left[E\left[\frac{I\{D=d\}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot E\left[\frac{S\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]}{\pi_0(d,X,Z)}\middle| D=d,X,Z\right]\middle| X,\Pi_0\right]\right]\\
= \quad & -E\left[E\left[\frac{I\{D=d\}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot\frac{\overbrace{E[S|D=d,X,Z]}^{=\pi_0(d,X,Z)}\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]}{\pi_0(d,X,Z)}\middle| X,\Pi_0\right]\right]\\
= \quad & -E\left[\frac{\overbrace{E[I\{D=d\}|X,\pi_0(d,X,Z)]}^{=p_{d0}(X,\pi_0(d,X,Z))}}{p_{d0}(X,\pi_0(d,X,Z))}\cdot[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))]\right]\\
= \quad & -E[\mu(d,1,X,\pi_0(d,X,Z)) - \mu_0(d,1,X,\pi_0(d,X,Z))],
\end{aligned}
$$

which cancels out with $(**)$. In an analogous way, it can be shown that $(***)$ corresponds to

$$
E[-\partial E[\mu_0(d,1,X,\pi_0(d,X,Z))]\cdot[\pi(d,X,Z)-\pi_0(d,X,Z)]],
$$

which cancels out with $(****)$. Therefore,

$$
\partial E\big[\phi_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big] = 0
$$

proving that the score function is orthogonal.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

This proof follows in a similar manner to the proof of theorem 1 and is omitted for brevity.

A.4 Proof of Theorem 4

Define the nuisance parameters to be the vector of functions $\eta = (p_d(X), \pi(D, X, M), \mu(D, S, X, M), \nu(D, S, X, M)$, with $p_d(X) = \Pr(D = d|X)$, $\pi(D, X, M) = \Pr(S = 1|D, X, M)$, $\mu(D, S, X, M) = E[Y|D, S, X, M]$, and $\nu(D, S, X, M) = \int E[Y|D, S, X, M = m]dF_{M=m|D,X}$, where $F_{M=m|D,X}$ denotes the conditional distribution function of M at value m. The score function for the counterfactual $\Psi_{d0} = E[Y(d)]$ is given by the following expression, with $W = (Y \cdot S, D, S, X, M)$:

$$\begin{aligned}\theta_d(W, \eta, \Psi_{d0}) &= \frac{I\{D = d\} \cdot S \cdot [Y - \mu(d, 1, X, M)]}{p_d(X) \cdot \pi(d, X, M)} \\ &+ \frac{I\{D = d\} \cdot [\mu(d, 1, X, M) - \nu(d, 1, X)]}{p_d(X)} \\ &+ \nu(d, 1, X) - \Psi_{d0}.\end{aligned}$$

Let $\mathcal{T}_n$ be the set fo all $\eta = (p_d, \pi, \mu, \nu)$ consisting of P-square integrable functions p_d, π, μ, and ν such that

$$\begin{aligned}\|\eta - \eta_0\|_q &\leq C, \qquad (A.8)\\ \|\eta - \eta_0\|_2 &\leq \delta_n, \\ \|p_d(X) - 1/2\|_\infty &\leq 1/2 - \epsilon, \\ \|\pi(D, X, M) - 1/2)\|_\infty &\leq 1/2 - \epsilon, \\ \|\mu(D, 1, X, M) - \mu_0(D, 1, X, M)\|_2 \times \|p_d(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}, \\ \|\mu(D, 1, X, M) - \mu_0(D, 1, X, M)\|_2 \times \|\pi(D, X, M) - \pi_0(D, X, M)\|_2 &\leq \delta_n n^{-1/2}, \\ \|\nu(d, 1, X) - \nu_0(d, 1, X)\|_2 \times \|p_d(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}.\end{aligned}$$

We furthermore replace the sequence $(\delta_n)_{n\geq 1}$ by $(\delta'_n)_{n\geq 1}$, where $\delta'_n = C_\epsilon \max(\delta_n, n^{-1/2})$, where C_ϵ is sufficiently large constant that only depends on C and ϵ.

Assumption 3.1: Linear scores and Neyman orthogonality

Assumption 3.1(a)

Moment Condition: The moment condition $E\Big[\theta_d(W,\eta_0,\Psi_{d0})\Big]=0$ is satisfied:

$$\begin{aligned}
E\Big[\theta_d(W,\eta_0,\Psi_{d0})\Big] &= E\Bigg[\overbrace{E\left[\frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)}\middle|X,M\right]}^{=E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M]|D=d,X]=0}\Bigg] \\
&+ E\Bigg[\overbrace{E\Big[\frac{I\{D=d\}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]}{p_{d0}(X)}\Big|X\Big]}^{=\int E\left[\mu_0(d,1,X,M)-\nu_0(d,1,X)\middle|D=d,X=x,M=m\right]dF_{M=m|D=d,X}=0}\Bigg] \\
&+ E\big[\nu_0(d,1,X)\big] - \Psi_{d0} = \Psi_{d0} - \Psi_{d0} = 0,
\end{aligned}$$

where the first equality follows from the law of iterated expectations. To better see this result, note that

$$\begin{aligned}
& E\left[\frac{I\{D=d\}\cdot S}{p_{d0}(X)\cdot\pi_0(d,X,M)}\cdot[Y-\mu_0(d,1,X,M)]\middle|X\right] \\
= \; & E\left[\frac{S}{\pi_0(d,X,M)}\cdot[Y-\mu_0(d,1,X,M)]\middle|D=d,X\right] \\
= \; & E\left[E\left[\frac{S}{\pi_0(d,X,M)}\cdot[Y-\mu_0(d,1,X,M)]\middle|D=d,X,M\right]\middle|D=d,X\right] \\
= \; & E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M|D=d,X]] \\
= \; & E[\mu_0(d,1,X,M)-\mu_0(d,1,X,M)|D=d,X]=0,
\end{aligned}$$

where the first and third equalities follow from basic probability theory and the second from the law of iterated expectations. Furthermore,

$$\begin{aligned}
& E\left[\frac{I\{D=d\}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]}{p_{d0}(X)}\middle|X=x\right] \\
= \; & E\big[\mu_0(d,1,X,M)-\nu_0(d,1,X)\big|D=d,X=x\big] \\
= \; & \int E\big[\mu_0(d,1,X,M)-\nu_0(d,1,X)\big|D=d,X=x,M=m\big]dF_{M=m|D=d,X=x} \\
= \; & \int E\big[\mu_0(d,1,X,M)\big|D=d,X=x,M=m\big]dF_{M=m|D=d,X=x}-\nu_0(d,1,x) \\
= \; & \nu(d,1,x)-\nu(d,1,x)=0.
\end{aligned}$$

where the first equality follows from basic probability theory, the second from conditioning on and integrating over M, and the third from the fact that $\nu_0(d,1,X)$ is not a function of M.

Assumption 3.1(b)

Linearity: The score $\theta_d(W,\eta_0,\Psi_{d0})$ is linear in Ψ_{d0} : $\theta_d(W,\eta_0,\Psi_{d0}) = \theta_d^a(W,\eta_0)\cdot\Psi_{d0}+\theta_d^b(W,\eta_0)$ with $\theta_d^a(W,\eta_0)=-1$ and

$$\begin{aligned}
\theta_d^b(W,\eta_0) &= \frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)} \\
&+ \frac{I\{D=d\}\cdot[\mu(d,1,X,M)-\nu(d,1,X)]}{p_d(X)}+\nu(d,1,X).
\end{aligned}$$

Assumption 3.1(c)

Continuity: The expression for the second Gateaux derivative of a map $\eta\mapsto E[\theta_d(W,\eta,\Psi_d)]$ is continuous.

Assumption 3.1(d)

Neyman Orthogonality: For any $\eta \in \mathcal{T}_N$, the Gateaux derivative in the direction $\eta - \eta_0 = (p_d(X) - p_{d0}(X), \pi(d,X,M) - \pi_0(d,X,M), \mu(D,1,X,M) - \mu_0(D,1,X,M), \nu(d,1,X) - \nu_0(d,1,X))$ is given by:

$$
\begin{aligned}
&\partial E\big[\theta_d(W,\eta,\Psi_d)\big]\big[\eta-\eta_0\big] = \\
&- E\left[\frac{I\{D=d\}\cdot S\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)}\right] && (*)\\
&+ E\left[\frac{I\{D=d\}\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{p_{d0}(X)}\right] && (**)\\
&\quad E[\cdot|X]=E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M]|D=d,X]=0 \\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,M)]}}{p_{d0}(X)\cdot\pi_0(d,X,M)}\cdot\frac{p_d(X)-p_{d0}(X)}{p_{d0}(X)}\right] \\
&\quad E[\cdot|X]=\int E\big[\mu_0(d,1,X,M)-\nu_0(d,1,X)\big|D=d,X=x,M=m\big]dF_{M=m|D=d,X=x}=0 \\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]}}{p_{d0}(X)}\cdot\frac{p_d(X)-p_{d0}(X)}{p_{d0}(X)}\right] \\
&\quad E[\cdot|X]=E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M]|D=d,X]=0 \\
&- E\left[\frac{\overbrace{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,M)]}}{p_{d0}(X)\cdot\pi_0(d,X,M)}\cdot\frac{\pi(d,X,M)-\pi_0(d,X,M)}{\pi_0(d,X,M)}\right] \\
&- \underbrace{E\left[\underbrace{\frac{I\{D=d\}\cdot[\nu(d,1,X)-\nu_0(d,1,X)]}{p_{d0}(X)}}_{E[\cdot|X]=\frac{p_{d0}(X)}{p_{d0}(X)}\cdot[\nu(d,1,X)-\nu_0(d,1,X)]}\right] + E[\nu(d,1,X)-\nu_0(d,1,X)]}_{=0} = 0
\end{aligned}
$$

The Gateaux derivative is zero because expressions $(*)$ and $(**)$ cancel out. To see this, note that

$$
\begin{aligned}
& E\left[\frac{I\{D=d\}\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{p_d(X)}\middle| X=x\right] \\
= \;& E\big[\mu(d,1,X,M)-\mu_0(d,1,X,M)\big|D=d,X=x\big] \\
= \;& \int E\big[\mu(d,1,X,M)-\mu_0(d,1,X,M)\big|D=d,X=x,M=m\big]dF_{M=m|D=d,X=x},
\end{aligned}
$$

where the first equality follows from basic probability theory and the second from conditioning on and integrating over M. Furthermore,

$$
\begin{aligned}
& E\left[\frac{I\{D=d\}\cdot S\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{p_d(X)\cdot\pi_0(d,X,M)}\middle| X=x\right] \\
= \;& E\left[\frac{S\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{\pi_0(d,X,M)}\middle| D=d,X=x\right] \\
= \;& \int E\left[\frac{S\cdot[\mu(d,1,X,M)-\mu_0(d,1,X,M)]}{\pi_0(d,X,M)}\middle| D=d,X=x,M=m\right]dF_{M=m|D=d,X=x} \\
= \;& \int E\big[\mu(d,1,X,M)-\mu_0(d,1,X,M)\big|D=d,S=1,X=x,M=m\big]dF_{M=m|D=d,X=x} \\
= \;& \int E\big[\mu(d,1,X,M)-\mu_0(d,1,X,M)\big|D=d,X=x,M=m\big]dF_{M=m|D=d,X=x},
\end{aligned}
$$

where the first equality follows from basic probability theory, the second from conditioning on and integrating

over M, the third from basic probability theory, and the fourth from simplification as $\mu(d,1,X,M) = E[Y|D = d, S = 1, X = x, M = m]$ is already conditional on $S = 1$.

$$\partial E\big[\theta_d(W,\eta,\Psi_d)\big]\big[\eta - \eta_0\big] = 0$$

proving that the score function is orthogonal.

Assumption 3.1(e)

Singular values of $E[\theta_d^a(W;\eta_0)]$ are bounded: This holds trivially, because $\theta_d^a(W;\eta_0) = -1$.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

Assumption 3.2(a)

This assumption directly follows from the construction of the set $\mathcal{T}_n$ and the regularity conditions (Assumption 12).

Assumption 3.2(b)

Bound for m_n:

$$\begin{aligned}
\|\mu_0(D,S,X,M)\|_q &= \left(E\left[|\mu_0(D,S,X,M)|^q\right]\right)^{\frac{1}{q}} \\
&= \left(\sum_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X,M)|^q \Pr_P(D=d,S=s|X,M)\right]\right)^{\frac{1}{q}} \\
&\geq \epsilon^{2/q}\left(\sum_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X,M)|^q\right]\right)^{\frac{1}{q}} \\
&\geq \epsilon^{2/q}\left(\max_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} E\left[|\mu_0(d,s,X,M)|^q\right]\right)^{\frac{1}{q}} \\
&= \epsilon^{2/q}\left(\max_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} \|\mu_0(d,s,X,M)\|_q\right),
\end{aligned}$$

where the first equality follows from definition, the second from the law of total probability, and the third line from the fact that $\Pr(D = d, S = 1|X,M) = p_{d0}(X)\cdot\pi_0(d,X,M) \geq \epsilon^2$ and $\Pr(D = d, S = 0|X,M) = p_{d0}(X)\cdot(1-\pi_0(d,X,M)) \geq \epsilon^2$. Similarly,

$$\|\nu_0(D,S,X)\|_q \geq \epsilon^{2/q}\left(\max_{d\in\{0,1,\ldots,Q\},s\in\{0,1\}} \|\nu_0(d,s,X)\|_q\right).$$

Notice that by Jensen's inequality $\|\mu_0(D,S,X,M)\|_q \leq \|Y\|_q$ and $\|\nu_0(D,S,X)\|_q \leq \|Y\|_q$ and hence $\|\mu_0(d,1,X,M)\|_q \leq C/\epsilon^{2/q}$ and $\|\nu_0(d,1,X)\|_q \leq C/\epsilon^{2/q}$, by conditions (A.8). Similarly, for any $\eta \in \mathcal{T}_N$: $\|\mu(d,1,X,M) - \mu_0(d,1,X,M)\|_q \leq C/\epsilon^{2/q}$ and $\|\nu(d,1,X) - \nu_0(d,1,X)\|_q \leq C/\epsilon^{2/q}$, because $\|\mu(D,S,X,M) - \mu_0(D,S,X,M)\|_q \leq C$ and $\|\nu(D,S,X) - \nu_0(D,S,X)\|_q \leq C$.

Consider

$$\begin{aligned} E\Big[\theta_d(W,\eta,\Psi_{d0})\Big] &= E\Bigg[\underbrace{\frac{I\{D=d\}\cdot S}{p_d(X)\cdot\pi(d,X,M)}\cdot Y}_{=I_1} \\ &+ \underbrace{\frac{I\{D=d\}}{p_d(X)}\cdot\left(1-\frac{S}{\pi(d,X,M)}\right)\cdot\mu(d,1,X,M)}_{=I_2} \\ &+ \underbrace{\left(1-\frac{I\{D=d\}}{p_d(X)}\right)\nu(d,1,X)}_{=I_3} - \Psi_{d0}\Bigg] \end{aligned}$$

and thus

$$\begin{aligned} \|\theta_d(W,\eta,\Psi_{d0})\|_q &\leq \|I_1\|_q + \|I_2\|_q + \|I_3\|_q + \|\Psi_{d0}\|_q \\ &\leq \frac{1}{\epsilon^2}\|Y\|_q + \frac{1-\epsilon}{\epsilon^2}\|\mu(d,1,X,M)\|_q + \\ &+ \frac{1-\epsilon}{\epsilon}\|\nu(d,1,X)\|_q + |\Psi_{d0}| \\ &\leq C\left(\frac{1}{\epsilon^2} + \frac{2(1-\epsilon)}{\epsilon^{2/q}}\left(\frac{1}{\epsilon^2}+\frac{1}{\epsilon}\right)+\frac{1}{\epsilon}\right), \end{aligned}$$

because of triangular inequality and because the following set of inequalities hold:

$$\begin{aligned} \|\mu(d,1,X,M)\|_q &\leq \|\mu(d,1,X,M)-\mu_0(d,1,X,M)\|_q + \|\mu_0(d,1,X,M)\|_q \leq 2C/\epsilon^{2/q}, \qquad \text{(A.9)} \\ \|\nu(d,1,X)\|_q &\leq \|\nu(d,1,X)-\nu_0(d,1,X)\|_q + \|\nu_0(d,1,X)\|_q \leq 2C/\epsilon^{2/q}, \\ |\Psi_{d0}| &= |E[\nu_0(d,1,X)]| \leq E\Big[\,|\nu_0(d,1,X)|^1\,\Big]^{\frac{1}{1}} = \|\nu_0(d,1,X)\|_1 \\ &\leq \|\nu_0(d,1,X)\|_2 \leq \|Y\|_2/\epsilon^{2/2} \overbrace{\leq}^{q>2} \|Y\|_q/\epsilon \leq C/\epsilon. \end{aligned}$$

which gives the upper bound on m_n in Assumption 3.2(b) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

Bound for m_n':

Notice that

$$\Big(E[|\theta_d^a(W,\eta)|^q]\Big)^{1/q} = 1$$

and this gives the upper bound on m_n' in Assumption 3.2(b).

Assumption 3.2(c)

In the following, we omit arguments for the sake of brevity and use $p_d = p_d(X), \pi = \pi(d,X,M), \nu = \nu(d,1,X), \mu = \mu(d,1,X,M)$ and similarly for $p_{d0}, \pi_0, \nu_0, \mu_0$.

Bound for r_n:

For any $\eta = (p_d, \pi, \mu, \nu)$ we have

$$\Big|E\Big(\theta_d^a(W,\eta) - \theta_d^a(W,\eta_0)\Big)\Big| = |1-1| = 0 \leq \delta_N',$$

and thus we have the bound on r_n from Assumption 3.2(c).

Bound for r_n':

$$\begin{aligned}
&\|\theta_d(W,\eta,\Psi_{d0})-\theta_d(W,\eta_0,\Psi_{d0})\|_2 \leq \left\|I\{D=d\}\cdot S\cdot Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2 \qquad \text{(A.10)}\\
+\quad &\left\|I\{D=d\}\cdot S\cdot\left(\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right)\right\|_2+\left\|I\{D=d\}\cdot\left(\frac{\mu}{p_d}-\frac{\mu_0}{p_{d0}}\right)\right\|_2\\
+\quad &\left\|I\{D=d\}\left(\frac{\nu}{p_d}-\frac{\nu_0}{p_{d0}}\right)\right\|_2+\|\nu-\nu_0\|_2\\
\leq\quad &\left\|Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2+\left\|\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right\|_2+\left\|\frac{\mu}{p_d}-\frac{\mu_0}{p_{d0}}\right\|_2+\left\|\frac{\nu}{p_d}-\frac{\nu_0}{p_{d0}}\right\|_2+\|\nu-\nu_0\|_2\\
\leq\quad &\frac{C}{\epsilon^4}\delta_n\left(1+\frac{1}{\epsilon}\right)+\delta_n\left(\frac{1}{\epsilon^5}+C+\frac{C}{\epsilon}\right)+\delta_n\left(\frac{1}{\epsilon^3}+\frac{C}{\epsilon^2}\right)+\delta_n\left(\frac{1}{\epsilon^3}+\frac{C}{\epsilon^2}\right)+\frac{\delta_n}{\epsilon}\leq\delta_n'
\end{aligned}$$

as long as C_ϵ in the definition of δ_n' is sufficiently large. This gives the bound on r_n' from Assumption 3.2(c). Here we made use of the fact that $\|\mu-\mu_0\|_2=\|\mu(d,1,X,M)-\mu_0(d,1,X,M)\|_2\leq\delta_n/\epsilon$, $\|\nu-\nu_0\|_2=\|\nu(d,1,X)-\nu_0(d,1,X)\|_2\leq\delta_n/\epsilon$ and $\|\pi-\pi_0\|_2=\|\pi(d,X)-\pi_0(d,X)\|_2\leq\delta_n/\epsilon$ using similar steps as in Assumption 3.1(b).

The last inequality in (A.10) is satisfied because we can bound the first term by

$$\begin{aligned}
&\left\|Y\cdot\left(\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right)\right\|_2\leq C\left\|\frac{1}{p_d\pi}-\frac{1}{p_{d0}\pi_0}\right\|_2\leq\frac{C}{\epsilon^4}\|p_{d0}\pi_0-p_d\pi\|_2\\
=\quad&\frac{C}{\epsilon^4}\|p_{d0}\pi_0-p_d\pi+p_{d0}\pi-p_{d0}\pi\|_2\leq\frac{C}{\epsilon^4}\left(\|p_{d0}(\pi_0-\pi)\|_2+\|\pi_0(p_{d0}-p_d)\|_2\right)\\
\leq\quad&\frac{C}{\epsilon^4}\left(\|\pi_0-\pi\|_2+\|p_{d0}-p_d\|_2\right)\leq\frac{C}{\epsilon^4}\delta_n\left(1+\frac{1}{\epsilon}\right),
\end{aligned}$$

where the first inequality follows from the second inequality in Assumption 4(a). The second term in (A.10) is bounded by

$$\begin{aligned}
&\left\|\frac{\mu}{p_d\pi}-\frac{\mu_0}{p_{d0}\pi_0}\right\|_2\leq\frac{1}{\epsilon^4}\|p_{d0}\pi_0\mu-p_d\pi\mu_0\|_2=\frac{1}{\epsilon^4}\|p_{d0}\pi_0\mu-p_d\pi\mu_0+p_{d0}\pi_0\mu_0-p_{d0}\pi_0\mu_0\|_2\\
\leq\quad&\frac{1}{\epsilon^4}\left(\|p_{d0}\pi_0(\mu-\mu_0)\|_2+\|\mu_0(p_{d0}\pi_0-p_d\pi)\|_2\right)\leq\frac{1}{\epsilon^4}\left(\|\mu-\mu_0\|_2+C\|p_{d0}\pi_0-p_d\pi\|_2\right)\\
\leq\quad&\frac{1}{\epsilon^4}\left(\frac{\delta_n}{\epsilon}+C\|p_{d0}\pi_0-p_d\pi\|_2\right)\leq\delta_n\left(\frac{1}{\epsilon^5}+C+\frac{C}{\epsilon}\right),
\end{aligned}$$

where the third inequality follows from $E[Y^2|D=d,S=s,X,M]\geq(E[Y|D=d,S=s,X,M])^2=\mu_0^2(d,s,X,M)$ by the conditional Jensen's inequality and therefore $\|\mu_0(d,s,X,M)\|_\infty\leq C^2$.

For the third term we get

$$\begin{aligned}
&\left\|\frac{\mu}{p_d}-\frac{\mu_0}{p_{d0}}\right\|_2=\frac{1}{\epsilon^2}\|p_{d0}\mu-p_d\mu_0\|_2=\frac{1}{\epsilon^2}\|p_{d0}\mu-p_d\mu_0+p_{d0}\mu_0-p_{d0}\mu_0\|_2\\
\leq\quad&\frac{1}{\epsilon^2}\left(\|p_{d0}(\mu-\mu_0)\|_2+\|\mu_0(p_{d0}-p_d)\|_2\right)\leq\frac{1}{\epsilon^2}\left(\|\mu-\mu_0\|_2+C\|p_{d0}-p_d\|_2\right)\leq\delta_n\left(\frac{1}{\epsilon^3}+\frac{C}{\epsilon^2}\right),
\end{aligned}$$

and similarly, for the fourth term we obtain

$$\left\|\frac{\nu}{p_d}-\frac{\nu_0}{p_{d0}}\right\|_2\leq\delta_n\left(\frac{1}{\epsilon^3}+\frac{C}{\epsilon^2}\right),$$

where we used Jensen's inequality twice to get $\|\nu_0(\underline{d}_2,X)\|_\infty\leq C^2$.

Bound for λ_n':

Now consider

$$f(r) := E[\theta(W; \Psi_{d0}, \eta + r(\eta - \eta_0)].$$

For any $r \in (0,1)$:

$$\begin{aligned}
\frac{\partial^2 f(r)}{\partial r^2} &= E\left[I\{D=d\} \cdot S \cdot (-2) \cdot \frac{(\mu-\mu_0)(p_d - p_{d0})}{(p_{d0} + r(p_d - p_{d0}))^2 (\pi_0 + r(\pi - \pi_0))}\right] \\
&+ E\left[I\{D=d\} \cdot S \cdot (-2) \cdot \frac{(\mu-\mu_0)(\pi - \pi_0)}{(p_{d0} + r(p_d - p_{d0})) (\pi_0 + r(\pi - \pi_0))^2}\right] \\
&+ E\left[I\{D=d\} \cdot S \cdot 2 \cdot \frac{(Y - \mu_0 - r(\mu-\mu_0))(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3 (\pi_0 + r(\pi - \pi_0))}\right] \\
&+ E\left[I\{D=d\} \cdot S \cdot 2 \cdot \frac{(Y - \mu_0 - r(\mu-\mu_0))(\pi - \pi_0)^2}{(p_{d0} + r(p_d - p_{d0})) (\pi_0 + r(\pi - \pi_0))^3}\right] \\
&+ E\left[I\{D=d\} \cdot S \cdot 2 \cdot \frac{(Y - \mu_0 - r(\mu-\mu_0))(p_d - p_{d0})(\pi - \pi_0)}{(p_{d0} + r(p_d - p_{d0}))^2 (\pi_0 + r(\pi - \pi_0))^2}\right] \\
&+ E\left[I\{D=d\} \cdot (-2) \cdot \frac{(\mu-\mu_0)(p_d - p_{d0})}{(p_{d0} + r(p_d - p_{d0}))^2}\right] + E\left[I\{D=d\} \cdot 2 \cdot \frac{(\nu-\nu_0)(p_d - p_{d0})}{(p_{d0} + r(p_d - p_{d0}))^2}\right] \\
&+ E\left[I\{D=d\} \cdot 2 \cdot \frac{r(\mu-\mu_0)(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3}\right] + E\left[I\{D=d\} \cdot 2 \cdot \frac{(r(\nu-\nu_0)(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3}\right] \\
&+ E\left[I\{D=d\} \cdot 2 \cdot \frac{(\mu_0-\nu_0)(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3}\right]
\end{aligned} \tag{A.11}$$

Note that because

$$\begin{aligned}
E[Y - \mu_0(d,1,X,M)|D=d, S=1, X, M] &= 0, \\
|p_d - p_{d0}| \le 2, \qquad |\pi - \pi_0| &\le 2 \\
\|\mu_0\|_q \le \|Y\|_q / \epsilon^{1/q} &\le C/\epsilon^{2/q} \\
\|\nu_0\|_q \le \|Y\|_q / \epsilon^{1/q} &\le C/\epsilon^{2/q} \\
\|\mu - \mu_0\|_2 \times \|p_d - p_{d0}\|_2 &\le \delta_n n^{-1/2}/\epsilon, \\
\|\mu - \mu_0\|_2 \times \|\pi - \pi_0\|_2 &\le \delta_n n^{-1/2}/\epsilon^2, \\
\|\nu - \nu_0\|_2 \times \|p_d - p_{d0}\|_2 &\le \delta_n n^{-1/2}/\epsilon.
\end{aligned}$$

we get that for some constant C''_ϵ that only depends on C and ϵ

$$\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \le C''_\epsilon \delta_n n^{-1/2} \le \delta'_n n^{-1/2}$$

and this gives the upper bound on λ'_n in Assumption 3.2(c) of Chernozhukov, Chetverikov, Demirer, Duflo,

Hansen, Newey, and Robins (2018) as long as $C_\epsilon \geq C''_\epsilon$. We used the following inequalities

$$\begin{aligned}
\|\mu - \mu_0\|_2 &= \|\mu(d,1,X,M) - \mu_0(d,1,X,M)\|_2 \leq \|\mu(D,S,X,M) - \mu_0(D,S,X,M)\|_2 / \epsilon \\
\|\nu - \nu_0\|_2 &= \|\nu(d,1,X) - \nu_0(d,1,X)\|_2 \leq \|\nu(D,S,X) - \nu_0(D,S,X)\|_2 / \epsilon^2 \\
\|\pi - \pi_0\|_2 &= \|\pi(d,X) - \pi_0(d,X)\|_2 \leq \|\pi(D,X) - \pi_0(D,X)\|_2 / \epsilon,
\end{aligned}$$

and these can be shown using similar steps as in Assumption 3.1(b).

To verify that $\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \leq C''_\epsilon \delta_n n^{-1/2}$ holds, note that by the triangular inequality it is sufficient to bound the absolute value of each of the ten terms in (A.11) separately. We illustrate it for the first, third, and last terms. For the first term:

$$\begin{aligned}
& \left| E\left[I\{D=d\} \cdot S(-2) \frac{(\mu-\mu_0)(p_d - p_{d0})}{(p_{d0} + r(p_d - p_{d0}))^2 (\pi_0 + r(\pi - \pi_0))} \right] \right| \\
\leq\ & 2 \left| E\left[\frac{(\mu-\mu_0)(p_d - p_{d0})}{(p_{d0} + r(p_d - p_{d0}))^2 (\pi_0 + r(\pi - \pi_0)} \right] \right| \\
\leq\ & \frac{2}{\epsilon^3} \left| E\left[(\mu-\mu_0)(p_d - p_{d0}) \right] \right| \leq \frac{2}{\epsilon^3} \frac{\delta_N}{\epsilon} n^{-1/2},
\end{aligned}$$

in the second inequality we used the fact that for $1 \geq p_{d0} + r(p_d - p_{d0}) = (1-r)p_{d0} + rp_d \geq (1-r)\epsilon + r\epsilon = \epsilon$ and similarly for π and in the third Holder's inequality. For the third term, we get

$$\begin{aligned}
& \left| E\left[I\{D=d\} \cdot S2 \frac{(Y - \mu_0 - r(\mu-\mu_0))(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3 (\pi_0 + r(\pi - \pi_0))} \right] \right| \\
\leq\ & \frac{2}{\epsilon^4} \left| E\left[I\{D=d\} \cdot S(Y - \mu_0 - r(\mu-\mu_0))(p_d - p_{d0})^2 \right] \right| \\
\leq\ & \frac{8}{\epsilon^4} \left| E\left[I\{D=d\} \cdot S(Y - \mu_0) \right] \right| + \frac{2}{\epsilon^4} \left| E\left[r(\mu-\mu_0)(p_d - p_{d0})^2 \right] \right| \\
\leq\ & \frac{2 \cdot 2}{\epsilon^4} \left| E\left[1 \cdot (\mu-\mu_0))(p_d - p_{d0}) \right] \right| \leq \frac{4}{\epsilon^4} \frac{\delta_N}{\epsilon} n^{-1/2},
\end{aligned}$$

where in addition we made use of conditions (A.8).

For the last term, we have

$$\begin{aligned}
& E\left[I\{D=d\} 2 \frac{(\mu_0 - \nu_0)(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3} \right] \\
&\qquad \int E\left[\mu_0(d,1,X,M) - \nu_0(d,1,X) \middle| D=d, X=x, M=m\right] dF_{M=m|D=d,X=x} = 0 \\
=\ & E\Bigg[\overbrace{I\{D=d\} \frac{(\mu_0 - \nu_0)}{p_{d0}}} \cdot \frac{2p_{d0}(p_d - p_{d0})^2}{(p_{d0} + r(p_d - p_{d0}))^3} \Bigg] = 0.
\end{aligned}$$

The remaining terms in (A.11) are bounded similarly.

Assumption 3.2(d)

$$
\begin{aligned}
E\Big[(\theta_d(W,\eta_0,\Psi_{d0}))^2\Big] &= E\Bigg[\Bigg(\underbrace{\frac{I\{D=d\}\cdot S\cdot[Y-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)}}_{=I_1} \\
&+ \underbrace{\frac{I\{D=d\}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]}{p_{d0}(X)}}_{=I_2} \\
&+ \underbrace{\nu_0(d,1,X)-\Psi_{d0}}_{=I_3}\Bigg)^2\Bigg] \\
&= E[I_1^2+I_2^2+I_3^2]\geq E[I_1^2] \\
&= E\left[I\{D=d\}\cdot S\cdot\left(\frac{[Y-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)}\right)^2\right] \\
&\geq \epsilon^2 E\left[\left(\frac{[Y-\mu_0(d,1,X,M)]}{p_{d0}(X)\cdot\pi_0(d,X,M)}\right)^2\right] \\
&\geq \frac{\epsilon^2 c^2}{(1-\epsilon)^4}>0,
\end{aligned}
$$

because $\Pr(D=d,S=1|X,M)=p_{d0}(X)\cdot\pi_0(d,X,M)\geq\epsilon^2,\ p_{d0}(X)\leq 1-\epsilon$ and $\pi_0(d,X,M)\leq 1-\epsilon$.

The the second equality follows from

$$
\begin{aligned}
E\Big[I_1\cdot I_2\Big] &= E\Bigg[\overbrace{\frac{I\{D=d\}\cdot S}{(p_{d0}(X))^2\cdot\pi_0(d,X,M)}\cdot[Y-\mu_0(d,1,X,M)]}^{E[\cdot|X]=E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M]|D=d,X]=0}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]\Bigg], \\
E\Big[I_2\cdot I_3\Big] &= E\Bigg[\overbrace{\frac{I\{D=d\}}{p_{d0}(X)}\cdot[\mu_0(d,1,X,M)-\nu_0(d,1,X)]}^{E[\cdot|X]=\int E\left[\mu_0(d,1,X,M)-\nu_0(d,1,X)\Big|D=d,X=x,M=m\right]dF_{M=m|D=d,X}=0}\cdot[\nu_0(d,1,X)-\Psi_{d0}]\Bigg], \\
E\Big[I_1\cdot I_3\Big] &= E\Bigg[\overbrace{\frac{I\{D=d\}\cdot S}{p_{d0}(X)\cdot\pi_0(d,X,M)}\cdot[Y-\mu_0(d,1,X,M)]}^{E[\cdot|X]=E[E[Y-\mu_0(d,1,X,M)|D=d,S=1,X,M]|D=d,X]=0}\cdot[\nu_0(d,1,X)-\Psi_{d0}]\Bigg].
\end{aligned}
$$

B Derivation of efficient influence functions

The proof that our estimators are based on efficient influence functions closely follows Levy (2019). For deriving the efficient influence functions under the identifying assumptions considered in Sections 3 and 2, let $\tilde{Y}=Y\cdot S$. Furthermore, denote the observed data by $O=(\tilde{Y}\cdot S,S,D,X)\sim P$, with distribution P having the density $f(o)=f_{\tilde{Y}}(y|d,s,x)f_{D,S}(d,s|x)f_X(x)$, where $f_A(a)$ is the unconditional density or probability of variable A at value a and $f_A(a|b)$ is the conditional density/probability given variable $B=b$.

Define $\Psi_{d0}=\Psi_d(P)=E_P[E_P[Y|D=d,S=1,X]]=E_P[E_P[\tilde{Y}|D=d,S=1,X]]$, where the second equality from the fact that $\tilde{Y}=Y$ for $S=1$. We now consider the derivative of $\Psi_{d^*}(P)$ w.r.t. the distribution P, with d^*

$\in \{0, 1, ..., Q\}$:

$$
\begin{aligned}
& \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \Psi_{d^*}(P_\epsilon) = E_{P_\epsilon}[E_{P_\epsilon}[\tilde{Y}|D=d^*, S=1, X]] \\
= & \int\int y \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} (f_{\tilde{Y},\epsilon}(y|d^*,1,x)dy f_{X,\epsilon}(x))dx \\
= & \int\int y \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} f_{\tilde{Y},\epsilon}(y|d^*,1,x)dy f_X(x)dx + \int\int y f_{\tilde{Y}}(y|d^*,1,x)dy \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} f_{X,\epsilon}(x)dx \\
= & \int\int\int y \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} f_{\tilde{Y},\epsilon}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \qquad \text{(B.1)} \\
+ & \int\int y f_{\tilde{Y}}(y|d^*,1,x)dy \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} f_{X,\epsilon}(x)dx. \qquad \text{(B.2)}
\end{aligned}
$$

Denote by $\mathcal{S} = \frac{\partial}{\partial \epsilon} \log f_\epsilon\Big|_{\epsilon=0}$ the score function (i.e. the derivative of the log density/likelihood). Applying identity (1) of Levy (2019), it follows that

$$
\begin{aligned}
\frac{\partial}{\partial \epsilon} f_{\tilde{Y},\epsilon}(y|d,s,x)|_{\epsilon=0} &= [\mathcal{S}(o) - E[\mathcal{S}(O)|d,s,x]] f_{\tilde{Y}}(y|d,s,x), \qquad \text{(B.3)} \\
\frac{\partial}{\partial \epsilon} f_{X,\epsilon}(x)|_{\epsilon=0} &= [E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x). \qquad \text{(B.4)}
\end{aligned}
$$

Plugging (B.3) and (B.4) into (B.1) and (B.2), respectively, yields

$$
\begin{aligned}
& \int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
- & \int\int\int y E[\mathcal{S}(O)|d,s,x] f_{\tilde{Y}}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
+ & \int\int y f_{\tilde{Y}}(y|d^*,1,x)dy [E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x)dx \\
= & \int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
- & \int\int E[\tilde{Y}|d,s,x] E[\mathcal{S}(O)|d,s,x] \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
+ & \int E[\tilde{Y}|d^*,1,x][E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x)dx \\
= & \int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
- & \int\int E[\tilde{Y}|d,s,x] \int \mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x)dy \frac{I\{D=d^*,S=1\}f_{D,S}(d,s|x)}{f_{D,S}(d,s|x)} d(d,s) f_X(x)dx \\
+ & \int E[\tilde{Y}|d^*,1,x] \int \mathcal{S}(o) f_{Y,\tilde{D},S}(y,d,s|x)d(y,d,s) f_X(x)dx - \int \mathcal{S}(o) f(o)do \int E[\tilde{Y}|d^*,1,x] f_X(x)dx \\
= & \int y\mathcal{S}(o) \frac{I\{D=d^*,S=1\}}{f_{D,S}(d,s|x)} f(o)do - \int E[\tilde{Y}|d,s,x]\mathcal{S}(o) \frac{I\{D=d^*,S=1\}}{f_{D,S}(d,s|x)} f(o)do \\
+ & \int E[\tilde{Y}|d^*,1,x]\mathcal{S}(o) f(o)do - \int \mathcal{S}(o)\Psi_{d^*}(P) f(o)do \\
= & \int \mathcal{S}(o) \left[\frac{I\{D=d^*,S=1\}}{f_{D,S}(d,s|x)} (y - E[\tilde{Y}|d,s,x]) + E[\tilde{Y}|d^*,1,x] - \Psi_{d^*}(P) \right] f(o)do. \qquad \text{(B.5)}
\end{aligned}
$$

(B.5) is an $L_0^2(P)$ inner product of the score $\mathcal{S}$ and the following function, which is thus the efficient influence

function:

$$\begin{aligned}
& \frac{I\{D=d^*,S=1\}}{f_{D,S}(d,s|x)}[Y-E[\tilde{Y}|D,S,X]]+E[\tilde{Y}|D=d^*,S=1,x]-\Psi_{d^*}(P) \\
= & \frac{I\{D=d^*,S=1\}\cdot[Y-E[Y|D=d^*,S=1,X]]}{f_{D,S}(d^*,1|x)}+E[Y|D=d^*,S=1,x]-\Psi_{d^*}(P) \\
= & \frac{I\{D=d^*\}\cdot S\cdot[Y-\mu(d^*,1,X)]}{p_{d^*}(X)\cdot\pi(d^*,X)}+\mu(d^*,1,X)-\Psi_{d^*}(P) \\
= & \psi_{d^*}-\Psi_{d^*}(P),
\end{aligned}$$

with ψ_{d^*} corresponding to (5) for $d^*=d$.

Analogously, one can define $\Psi_{d0}=\Psi_d(P)=E_P[E_P[Y|D=d,S=1,X,\Pi]]$ and show that the efficient influence function corresponds to $\phi_d-\Psi_d(P)$, with ϕ_d defined in (12). This follows straightforwardly from replacing X by X,Π everywhere in the previous derivations. In a similar manner, one can demonstrate that for $\Psi_d^{S=1}(P)=E_P[E_P[Y|D=d,S=1,X,\Pi]|S=1]$, the efficient influence function corresponds to $\phi_{d,S=1}-\Psi_d^{S=1}(P)$, with $\phi_{d,S=1}$ defined in (9). This follows from considering the derivative in the selected population with $S=1$ (rather than the total population) and replacing D,S by D as well as X by X,Π everywhere in the previous derivations. The proofs for these cases are thus omitted for the sake of brevity.

For deriving the efficient influence function under the identifying assumptions of Section 4, let $\tilde{Y}=Y\cdot S$ and denote the observed data by $O=(\tilde{Y}\cdot S,S,D,X,M)\sim P$, with distribution P having the density $f(o)=f_{\tilde{Y}}(y|d,s,x,m)f_S(s|d,x,m)f_M(m|d,x)f_D(d|x)f_X(x)$.

Define $\Psi_{d0}=\Psi_d(P)=E_P[E_P[E_P[Y|D=d,S=1,X,M]|D=d,X]]=E_P[E_P[E_P[\tilde{Y}|D=d,S=1,X,M]|D=d,X]]$, where the second equality from the fact that $\tilde{Y}=Y$ for $S=1$. We now consider the derivative of $\Psi_{d^*}(P)$ w.r.t. the distribution P, with $d^*\in\{0,1,...,Q\}$:

$$\begin{aligned}
& \frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}\Psi_{d^*}(P_\epsilon)=E_{P_\epsilon}[E_{P_\epsilon}[E_{P_\epsilon}[\tilde{Y}|D=d^*,S=1,X,M]|D=d^*,X]] \\
= & \int\int\int y\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}(f_{\tilde{Y},\epsilon}(y|d^*,1,x,m)dyf_{M,\epsilon}(m|d,x)dmf_{X,\epsilon}(x))dx \\
= & \int\int\int y\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}f_{\tilde{Y},\epsilon}(y|d^*,1,x)dyf_M(m|d,x)dmf_X(x)dx \\
+ & \int\int\int yf_{\tilde{Y}}(y|d^*,1,x,m)dy\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}f_{M,\epsilon}(m|d,x)dmf_X(x)dx \\
+ & \int\int\int yf_{\tilde{Y}}(y|d^*,1,x,m)dyf_M(m|d,x)dm\frac{\partial}{\partial\epsilon}dm\Big|_{\epsilon=0}f_{X,\epsilon}(x)dx \\
= & \int\int\int\int\int y\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}f_{\tilde{Y},\epsilon}(y|d,s,x,m)dy\frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)}dsf_M(m|d,x)dm\frac{f_D(d|x)}{f_D(d|x)}ddf_X(x)dx \qquad \text{(B.6)} \\
+ & \int\int\int\int yf_{\tilde{Y}}(y|d,1,x,m)dy\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=0}f_{M,\epsilon}(m|d,x)dm\frac{I\{D=d^*\}f_D(d|x)}{f_D(d|x)}ddf_X(x)dx \qquad \text{(B.7)} \\
+ & \int\int\int yf_{\tilde{Y}}(y|d^*,1,x,m)dyf_M(m|d,x)dm\frac{\partial}{\partial\epsilon}dm\Big|_{\epsilon=0}f_{X,\epsilon}(x)dx. \qquad \text{(B.8)}
\end{aligned}$$

Denote by $\mathcal{S}=\frac{\partial}{\partial\epsilon}\log f_\epsilon\Big|_{\epsilon=0}$ the score function (i.e. the derivative of the log density/likelihood). Applying

identity (1) of Levy (2019), it follows that

$$\frac{\partial}{\partial\epsilon} f_{\tilde{Y},\epsilon}(y|d,s,x,m)|_{\epsilon=0} = [\mathcal{S}(o) - E[\mathcal{S}(O)|d,s,x,m]] f_{\tilde{Y}}(y|d,s,x,m), \tag{B.9}$$

$$\frac{\partial}{\partial\epsilon} f_{M,\epsilon}(m|d,x)|_{\epsilon=0} = [E[\mathcal{S}(O)|m,d,x] - E[\mathcal{S}(O)|d,x]] f_M(m|d,x), \tag{B.10}$$

$$\frac{\partial}{\partial\epsilon} f_{X,\epsilon}(x)|_{\epsilon=0} = [E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x). \tag{B.11}$$

Plugging (B.9), (B.10), and (B.11) into (B.6), (B.7), and (B.8), respectively, yields

$$\begin{aligned}
& \int\int\int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x,m)dy \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
- & \int\int\int\int\int y E[\mathcal{S}(O)|d,s,x,m] f_{\tilde{Y}}(y|d,s,x,m)dy \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int\int\int y f_{\tilde{Y}}(y|d,1,x,m)dy [E[\mathcal{S}(O)|m,d,x] - E[\mathcal{S}(O)|d,x]] f_M(m|d,x)dm \frac{I\{D=d^*\}f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int\int y f_{\tilde{Y}}(y|d^*,1,x,m)dy f_M(m|d,x)dm [E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x)dx \\
= & \int\int\int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x,m)dy \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
- & \int\int\int\int E[\tilde{Y}|d,s,x,m] E[\mathcal{S}(O)|d,s,x,m] \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int\int E[\tilde{Y}|d,1,x,m] [E[\mathcal{S}(O)|m,d,x] - E[\mathcal{S}(O)|d,x]] f_M(m|d,x)dm \frac{I\{D=d^*\}f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int E[\tilde{Y}|d^*,1,x,m] f_M(m|d,x)dm [E[\mathcal{S}(O)|x] - E[\mathcal{S}(O)]] f_X(x)dx \\
= & \int\int\int\int\int y\mathcal{S}(o) f_{\tilde{Y}}(y|d,s,x,m)dy \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
- & \int\int\int\int E[\tilde{Y}|d,s,x,m] \int S(o) f_{\tilde{Y}}(y|d,s,x,m)dy \frac{I\{D=d^*,S=1\}f_S(s|d,x,m)}{f_S(s|d,x,m)} ds f_M(m|d,x)dm \frac{f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int\int E[\tilde{Y}|d,1,x,m] \int S(o) f_{\tilde{Y}}(y,s|d,x,m)d(y,s) f_M(m|d,x)dm \frac{I\{D=d^*\}f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
- & \int\int\int E[\tilde{Y}|d,1,x,m] f_M(m|d,x)dm \int S(o) f_{\tilde{Y}}(y,s,m|d,x)d(y,s,m) \frac{I\{D=d^*\}f_D(d|x)}{f_D(d|x)} dd f_X(x)dx \\
+ & \int\int E[\tilde{Y}|d^*,1,x,m] f_M(m|d,x)dm \int S(o) f_{\tilde{Y}}(y,d,s,m|x)d(y,d,s,m) f_X(x)dx \\
- & \int \mathcal{S}(o) f(o)do \int\int E[\tilde{Y}|d^*,1,x,m] f_M(m|d,x)dm f_X(x)dx \\
= & \int y\mathcal{S}(o) \frac{I\{D=d^*,S=1\}}{f_S(s|d,x,m)\cdot f_D(d|x)} f(o)do - \int E[\tilde{Y}|d,s,x,m] \frac{I\{D=d^*,S=1\}}{f_S(s|d,x,m)\cdot f_D(d|x)} f(o)do \\
+ & \int E[\tilde{Y}|d,1,x,m] S(o) \frac{I\{D=d^*\}}{f_D(d|x)} f(o)do - \int E[E[\tilde{Y}|d,1,x,m]|d,x] S(o) \frac{I\{D=d^*\}}{f_D(d|x)} f(o)do \\
+ & \int E[E[\tilde{Y}|d^*,1,x,m]|d^*,x] S(o) f(o)do - \int \mathcal{S}(o)\Psi_{d^*}(P) f(o)do \\
= & \int \mathcal{S}(o) \left[\frac{I\{D=d^*,S=1\}}{f_S(s|d,x,m)\cdot f_D(d|x)} (y - E[\tilde{Y}|d,s,x,m]) \right. \\
+ & \left. \frac{I\{D=d^*\}}{f_D(d|x)} (E[\tilde{Y}|d,1,x,m] - E[E[\tilde{Y}|d,1,x]|d,x]) + E[E[\tilde{Y}|d^*,1,x]|d^*,x] - \Psi_{d^*}(P) \right] f(o)do.
\end{aligned} \tag{B.12}$$

(B.12) is an $L_0^2(P)$ inner product of the score $\mathcal{S}$ and the following function, which is thus the efficient influence

function:

$$
\begin{aligned}
& \frac{I\{D=d^*, S=1\}}{f_S(s|D,X,M)\cdot f_D(d|X)}(y - E[\tilde{Y}|D,S,X,X]) \\
+ \quad & \frac{I\{D=d^*\}}{f_D(d|X)}(E[\tilde{Y}|D,S=1,X,M] - E[E[\tilde{Y}|D,S=1,X]|D,X]) + E[E[\tilde{Y}|D=d^*,S=1,X]|D=d^*,X] - \Psi_{d^*}(P) \\
= \quad & \frac{I\{D=d^*\}\cdot S\cdot[Y-\mu(d^*,1,X,M)]}{p_{d^*}(X)\cdot\pi(d^*,X,M)} + \frac{I\{D=d^*\}\cdot[\mu(d^*,1,X,M)-\nu(d^*,1,X)]}{p_{d^*}(X)} + \nu(d^*,1,X)] - \Psi_{d^*}(P) \\
= \quad & \psi_{d^*} - \Psi_{d^*}(P),
\end{aligned}
$$

with ψ_{d^*} corresponding to (15) for $d^* = d$.

C Descriptive Statistics

In the following tables we report descriptive statistics, namely the number of observations and means by treatment groups, for a selected set of pre-treatment covariates X measured at Job Corps assignment and post-treatment covariates M measured in the second and third year after assignment.[7]

[7] Additional descriptive statistics are available upon request from the authors.

variable name (X)	N	any train	N	no train	N	acad.	N	voc.
no child at random assignment	1,673	0.699	200	0.600	830	0.700	843	0.698
pregnant at random assignment	1,673	0.008	200	0.015	830	0.007	843	0.009
sample member does not contribute to rent	1,673	0.686	200	0.640	830	0.699	843	0.674
diploma at random assignment								
no HS diploma	1,673	0.757	200	0.755	830	0.853	843	0.662
no GED	1,673	0.954	200	0.960	830	0.970	843	0.938
no other degree	1,673	0.980	200	0.965	830	0.980	843	0.966
job and training in the past year								
no training	1,673	0.325	200	0.400	830	0.301	843	0.348
full time or part Time	1,673	0.230	200	0.210	830	0.270	843	0.190
no job	1,673	0.393	200	0.355	830	0.430	843	0.357
stayed in most recent job	1,673	0.188	200	0.190	830	0.184	843	0.192
public assistance in the past year								
no public assistance	1,673	0.302	200	0.250	830	0.278	843	0.326
no AFDC	1,673	0.536	200	0.500	830	0.520	843	0.550
no other welfare	1,673	0.673	200	0.700	830	0.645	843	0.701
no food stamps	1,673	0.445	200	0.375	830	0.413	843	0.476
did not have health problems that limited work	1,673	0.944	200	0.920	830	0.940	843	0.948
use of drugs and alcohol in the past year								
no use of alcohol	1,673	0.470	200	0.395	830	0.522	843	0.419
no use of marijuana	1,673	0.671	200	0.700	830	0.671	843	0.671
no use of cocaine	1,673	0.962	200	0.965	830	0.964	843	0.960
no use of crack	1,673	0.980	200	0.985	830	0.977	843	0.982
no use of heroin	1,673	0.985	200	0.990	830	0.981	843	0.989
arrested or convicted in past years (at least once)								
arrested	1,673	0.836	200	0.830	830	0.810	843	0.861
convicted for a crime	1,673	0.895	200	1.000	830	0.964	843	0.985
convicted for murder or Assault	1,673	0.952	200	0.915	830	0.881	843	0.910
convicted for robbery	1,673	0.959	200	0.915	830	0.881	843	0.910
convicted for burglary	1,673	0.958	200	0.915	830	0.881	843	0.910
convicted for larceny	1,673	0.920	200	0.965	830	0.936	843	0.967
convicted for drug violation	1,673	0.959	200	0.970	830	0.948	843	0.970
no arrest charges pending at random assignment	1,673	0.969	200	0.970	830	0.946	843	0.973

variable name (X)	N	any train	N	no train	N	acad.	N	voc.
reason for joining JC program								
to get away from home	1,673	0.611	200	0.585	830	0.623	843	0.600
to get away from community	1,673	0.644	200	0.605	830	0.710	843	0.580
to be trained	1,673	0.987	200	0.980	830	0.987	843	0.988
for career	1,673	0.995	200	1.000	830	0.994	843	0.996
to get GED	1,673	0.962	200	0.965	830	0.039	843	0.963
because unemployed	1,673	0.929	200	0.915	830	0.929	843	0.929
other reason	1,673	0.759	200	0.745	830	0.761	843	0.756
expectations about JC program								
JC improves math	1,673	0.745	200	0.770	830	0.801	843	0.690
JC improves reading	1,673	0.586	200	0.595	830	0.649	843	0.523
JC improves network	1,673	0.640	200	0.645	830	0.664	843	0.616
JC improves self control	1,673	0.595	200	0.595	830	0.625	843	0.566
JC improves self esteem	1,673	0.634	200	0.605	830	0.655	843	0.612
JC expected to train for specific jobs	1,673	0.957	200	0.970	830	0.954	843	0.960
JC leads to new friendship	1,673	0.715	200	0.690	830	0.700	843	0.706
training received last week before random assignment	1,673	0.015	200	0.015	830	0.724	843	0.012
worked last week before random assignment	1,673	0.209	200	0.215	830	0.018	843	0.205
no welfare receipt								
month1 - year before random assignment	1,673	0.345	200	0.280	830	0.313	843	0.376
month2 - year before random assignment	1,673	0.340	200	0.28	830	0.313	843	0.375
month3 - year before random assignment	1,673	0.345	200	0.280	830	0.310	843	0.377
month4 - year before random assignment	1,673	0.348	200	0.280	830	0.312	843	0.383
month5 - year before random assignment	1,673	0.344	200	0.280	830	0.308	843	0.378
month6 - year before random assignment	1,673	0.343	200	0.275	830	0.312	843	0.374
month7 - year before random assignment	1,673	0.340	200	0.275	830	0.307	843	0.372
month8 - year before random assignment	1,673	0.338	200	0.270	830	0.310	843	0.367
month9 - year before random assignment	1,673	0.337	200	0.270	830	0.310	843	0.363
month10 - year before random assignment	1,673	0.338	200	0.270	830	0.311	843	0.365
month11 - year before random assignment	1,673	0.340	200	0.275	830	0.307	843	0.371
month12 - year before random assignment	1,673	0.340	200	0.290	830	0.311	843	0.368
had a job at random assignment	1,673	0.791	200	0.785	830	0.787	843	0.795
support received by friends and parents for attending JC								
encouraged by parents to attend JC	1,673	0.983	200	0.965	830	0.980	843	0.986
encouraged by relatives to attend JC	1,673	0.981	200	0.975	830	0.981	843	0.982
encouraged by friends to attend JC	1,673	0.956	200	0.965	830	0.947	843	0.964
encouraged by teacher to attend JC	1,673	0.991	200	0.985	830	0.988	843	0.994
encouraged by case worker to attend JC	1,673	0.999	200	0.995	830	0.998	843	1.000
encouraged by officer to attend JC	1,673	0.999	200	1.000	830	1.000	843	0.999

variable name (M)	N	any train	N	no train	N	acad.	N	voc.
did not get unemployment benefits								
week1 (1 year after assignment)	1,673	0.997	200	1.000	830	0.996	843	0.998
week8	1,673	0.998	200	0.995	830	0.999	843	0.998
week18	1,673	0.998	200	0.995	830	0.999	843	0.998
week28	1,673	0.997	200	0.995	830	0.999	843	0.996
week38	1,673	0.998	200	0.995	830	0.999	843	0.999
week48	1,673	0.998	200	0.995	830	0.996	843	1.000
week52	1,673	0.999	200	0.995	830	0.999	843	1.000
not in JC								
week1 (1 year after assignment)	1,673	0.860	200	0.830	830	0.847	843	0.873
week8	1,673	0.946	200	0.835	830	0.931	843	0.962
week18	1,673	0.931	200	0.785	830	0.905	843	0.957
week28	1,673	0.910	200	0.750	830	0.867	843	0.951
week38	1,673	0.899	200	0.780	830	0.866	843	0.931
week48	1,673	0.884	200	0.800	830	0.852	843	0.916
week52	1,673	0.880	200	0.785	830	0.853	843	0.906
not in a drug treatment program								
week1 (1 year after assignment)	1,673	0.999	200	1.000	830	1.000	843	0.998
week8	1,673	0.997	200	0.995	830	0.997	843	0.997
week18	1,673	0.996	200	1.000	830	0.996	843	0.996
week28	1,673	0.997	200	1.000	830	0.997	843	0.996
week38	1,673	0.997	200	1.000	830	0.996	843	0.998
week48	1,673	0.995	200	1.000	830	0.993	843	0.997
week52	1,673	0.994	200	1.000	830	0.991	843	0.996

variable name (M)	N	any train	N	no train	N	acad.	N	voc.
earnings								
week53 (2 years after assignment)	1,673	73.875	200	70.782	830	64.641	843	82.965
week60	1,673	87.341	200	83.211	830	74.509	843	99.974
week70	1,673	100.763	200	97.380	830	86.827	843	114.484
week80	1,673	114.549	200	95.989	830	99.868	843	129.003
week90	1,673	127.213	200	93.454	830	104.594	843	149.483
week100	1,673	131.921	200	100.158	830	111.555	843	151.973
week104	1,673	136.584	200	102.975	830	115.530	843	157.313
hours worked								
week53 (2 years after assignment)	1,673	12.952	200	12.778	830	11.592	843	14.291
week60	1,673	14.339	200	14.600	830	12.388	843	16.259
week70	1,673	15.764	200	15.646	830	13.742	843	17.755
week80	1,673	17.552	200	15.450	830	15.698	843	19.378
week90	1,673	19.345	200	16.177	830	16.319	843	22.325
week100	1,673	20.145	200	16.925	830	17.596	843	22.654
week104	1,673	20.699	200	17.418	830	18.248	843	23.112
employed								
week53 (2 years after assignment)	1,673	0.341	200	0.335	830	0.308	843	0.374
week60	1,673	0.354	200	0.350	830	0.304	843	0.403
week70	1,673	0.385	200	0.415	830	0.337	843	0.432
week80	1,673	0.429	200	0.380	830	0.380	843	0.476
week90	1,673	0.468	200	0.400	830	0.395	843	0.539
week100	1,673	0.481	200	0.430	830	0.413	843	0.548
week104	1,673	0.497	200	0.440	830	0.432	843	0.561