

# Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting

Yu-Chin Hsu<sup>†</sup>

Institute of Economics, Academia Sinica

Martin Huber<sup>‡</sup>

Department of Economics, University of Fribourg

Tsung-Chih Lai<sup>§</sup>

Department of Economics, Feng Chia University

**Abstract:** Using a sequential conditional independence assumption, this paper discusses fully nonparametric estimation of natural direct and indirect causal effects in causal mediation analysis based on inverse probability weighting. We propose estimators of the average indirect effect of a binary treatment, which operates through intermediate variables (or mediators) on the causal path between the treatment and the outcome, as well as the unmediated direct effect. In a first step, treatment propensity scores given the mediator and observed covariates or given covariates alone are estimated by nonparametric series logit estimation. In a second step, they are used to reweigh observations in order to estimate the effects of interest. We establish root-n consistency and asymptotic normality of this approach as well as a weighted version thereof. The latter allows evaluating effects on specific subgroups like the treated, for which we derive the asymptotic properties under estimated propensity scores. We also provide a simulation study and an application to an information intervention about male circumcisions.

**Keywords:** Causal mechanisms, direct effects, indirect effects, causal channels, mediation analysis, causal pathways, series logit estimation, nonparametric estimation, inverse probability weighting, propensity score.

**JEL classification:** C21.

# 1 Introduction

While a large literature in social sciences focusses on assessing the average treatment effect (ATE) of some intervention, quite frequently, also the causal mechanisms through which the effect materializes appear interesting. Gelman and Imbens (2013), for instance, argue that in many cases not only the ‘effects of causes’ seem relevant, but also the ‘causes of effects’. When for example assessing the earnings effect of a training program, policy makers might want to know whether the total impact comes from a change in search effort, human capital, or other mediators that are themselves affected by the training. For this reason, causal mediation analysis aims at disentangling a treatment effect into the indirect effect operating through one or several mediators as well as the direct effect, net of mediation. Even under random treatment assignment, total effects can in general not be disentangled by bluntly controlling for mediators, because this likely introduces selection bias, see Robins and Greenland (1992). However, direct and indirect effects are identified under a particular sequential conditional independence assumption that assumes the exogeneity of the treatment given observed covariates and of the mediator given observed covariates and the treatment, see for instance Imai, Keele, and Yamamoto (2010). Huber (2014) shows that under this assumption, identification is obtained by weighting observations by the inverses of particular treatment propensity scores<sup>1</sup> and considers semiparametric estimation based on parametric propensity score models in a simulation study and an application.

This paper is the first to consider fully nonparametric estimation of natural direct and indirect effects (in the denomination of Pearl (2001)) based on inverse probability weighting (IPW), using series logit estimation for the computation of the propensity scores. The advantage of the latter approach is that it prevents inconsistency of IPW due to an incorrectly specified parametric functional form of the propensity scores. We formally show that under the sequential conditional independence assumption and particular regularity conditions, nonparametric IPW is root-n consistent and asymptotically normal. Furthermore, our estimator attains the semiparametric efficiency bounds for mediation analysis derived by Tchetgen Tchetgen and Shpitser (2012). We therefore contribute to a growing literature concerned with assessing direct and indirect effects based on conditional independence under rather flexible model

---

<sup>1</sup>Tchetgen Tchetgen (2013) derives a related result in the context of inverse odds-ratio weighting.

assumptions,<sup>2</sup> see for instance Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), Flores and Flores-Lagunes (2009), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), and Vansteelandt, Bekaert, and Lange (2012), among others. In addition to the evaluation of these effects in the total population, we in contrast to Huber (2014) also discuss the identification and estimation of weighted direct and indirect effects. This provides a framework for evaluating causal parameters in interesting subgroups, such as the direct and indirect effects on the treated, which are explicitly considered in this paper. Also for the estimators of the weighted effects in general and the effects on the treated with estimated propensity scores in particular, we show root-n consistency and asymptotic normality.

Furthermore, we investigate the finite sample performance of nonparametric IPW in a simulation study and compare it to other estimators considered in the literature, namely maximum-likelihood-based g-computation (suggested by Robins (1986) and considered in the context of direct and indirect effects in Zheng and van der Laan (2012)), estimation based on (parametrically) simulating potential mediators and outcomes using the ‘mediation’ package for R by Tingley, Yamamoto, Hirose, Imai, and Keele (2014), ‘multiply robust’ estimation based on the efficient influence function as suggested by Tchetgen Tchetgen and Shpitser (2012), and IPW with parametric propensity scores as in Huber (2014). Finally, we apply our estimator (as well as the other methods considered in the simulations) to the experimental evaluation of Chinkhumba, Godlonton, and Thornton (2014) who investigate an information intervention about male circumcisions and HIV risk in urban Malawi. We investigate whether the information treatment affects the event/willingness of being circumcised indirectly through a change in the assessment of the relative HIV risk for circumcised and uncircumcised males (which serves as mediator), or ‘directly’, i.e. through other mechanisms. The results point to a small but quite robust indirect effect, while the direct effect estimates are never statistically different from zero.

The remainder of this paper is organized as follows. Section 2 defines the average natural direct and indirect effects and presents the identifying assumptions. Section 3 discusses nonparametric estimation based on IPW as well as inference and shows root-n consistency and asymptotic normality under particular regularity conditions. Sections 4 extends the identification and

---

<sup>2</sup>In contrast, the seminal papers in mediation analysis of Judd and Kenny (1981) and Baron and Kenny (1986) assume linear models for both the mediator and the outcome.

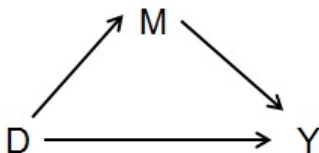
estimation results to weighted effects and to the effects on the treated under estimated propensity scores, respectively. Section 5 provides a simulation study, while Section 6 presents an application in the field of development economics. Section 7 concludes. All proofs are deferred to the Appendix.

## 2 Parameters of interest and identifying assumptions

### 2.1 Natural direct and indirect effects

We denote by  $D$  a binary intervention or treatment variable and by  $Y$  the outcome variable of interest. We would like to disentangle the causal effect of  $D$  on  $Y$  into a direct effect and an indirect impact that works through one or several discrete or continuous intermediate variables or mediators, denoted by  $M$ . Figure 1 provides a graphical illustration of the causal framework, in which arrows represent causal effects from one variable to another, but any confounders are omitted for the sake of simplicity. The definition of the total, direct, and indirect effects makes use of the potential outcome framework, see for instance Rubin (1974), and its adaptation to mediation analysis, see for instance Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008).  $M(d)$ ,  $Y(d, M(d))$  denote the potential mediator state and the potential outcome, respectively, under treatment  $d \in \{0, 1\}$ .  $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$  yields the (total) average causal effect, also known as average treatment effect (ATE), which has received much attention in the treatment or program evaluation literature, see for instance Imbens and Wooldridge (2009) for a survey.

Figure 1: Graphical illustration of the mediation framework



The (average) natural direct effect (using the denomination of Pearl (2001))<sup>3</sup> is defined as the mean effect of varying the treatment when keeping the mediator fixed at its potential value for

---

<sup>3</sup>Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure direct effect and Flores and Flores-Lagunes (2009) as net average treatment effect.

some  $d \in \{0, 1\}$ :

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\}. \quad (1)$$

Analogously, the (average) indirect effect is defined as the mean effect of shifting the mediator to its potential values under treatment and non-treatment when keeping the treatment fixed:

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\}. \quad (2)$$

The ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

$$\begin{aligned} \Delta &= E[Y(1, M(1)) - Y(0, M(0))] \\ &= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \theta(0) + \delta(1) \\ &= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \theta(1) + \delta(0). \end{aligned} \quad (3)$$

The notation  $\theta(1), \theta(0)$  and  $\delta(1), \delta(0)$  point out that direct and indirect effects may be heterogeneous in the treatment, which allows for interaction effects between the treatment and the mediator. No effect is identifiable without assumptions, because either  $Y(1, M(1))$  or  $Y(0, M(0))$  (but never both) is known for any individual, while  $Y(1, M(0))$  and  $Y(0, M(1))$  cannot be observed for anyone (as an individual can either be treated or non-treated, but not both at the same time).

## 2.2 Identification

Like Imai, Keele, and Yamamoto (2010) and many others, we rely on a sequential conditional independence assumption for identification. To this end, let  $X$  denote a vector of observed covariates that potentially confound the treatment, the mediator, and the outcome. Furthermore, we denote by  $\mathcal{X}, \mathcal{M}$  the supports of  $X$  and  $M$ , respectively.

**Assumption 1 (conditional independence of the treatment):**

$$\{Y(d', m), M(d)\} \perp D | X = x \text{ for all } d', d \in \{0, 1\} \text{ and } m, x \text{ in } \mathcal{M} \times \mathcal{X}.$$

Assumption 1 requires the treatment to be conditionally independent of the potential mediator states and outcomes given  $X$ , ruling out unobserved confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on the

covariates. This restriction is known as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature.

**Assumption 2 (conditional independence of the mediator):**

$Y(d', m) \perp M | D = d, X = x$  for all  $d', d \in \{0, 1\}$  and  $m, x$  in  $\mathcal{M} \times \mathcal{X}$ .

Assumption 2 requires the mediator to be conditionally independent of the potential outcomes given  $D$  and  $X$ , ruling out unobserved confounders jointly causing the mediator and the outcome conditional on the treatment and the covariates.

**Assumption 3 (common support):**

$\Pr(D = d | M = m, X = x)$  is bounded away from zero for all  $d \in \{0, 1\}$  and  $m, x$  in  $\mathcal{M} \times \mathcal{X}$ .

Assumption 3 is a common support restriction requiring that the conditional probability to be treated given  $M, X$  must be bounded away from zero in either treatment state. Note that  $\Pr(D = d | X = x)$  must therefore be bounded away from zero on  $\mathcal{X}$ , too. By Bayes' theorem, Assumption 3 also implies  $\Pr(M = m | D = d, X = x) > 0$  or, in the case of a continuous  $M$ , that the conditional density of  $M$  given  $D, X$  is larger than zero. Therefore, the mediator state must not be a deterministic function of the treatment conditional on  $X$ .

Identification of the natural direct and indirect effects under these or similar assumptions based on functions of the conditional mean of  $Y$  given  $D, M, X$  and the conditional density of  $M$  given  $D, X$  (the mediation formulae) has been demonstrated for instance in Pearl (2001) and Imai, Keele, and Yamamoto (2010). Huber (2014) shows that the effects as well as the mean potential outcomes may alternatively be identified by inverse probability weighting (IPW) based on the conditional probabilities  $\Pr(D = 1 | M, X)$  and  $\Pr(D = d | X)$ , henceforth referred to as propensity scores:

$$\begin{aligned} \theta(d) &= E \left[ \left( \frac{YD}{\Pr(D = 1 | M, X)} - \frac{Y(1 - D)}{1 - \Pr(D = 1 | M, X)} \right) \frac{\Pr(D = d | M, X)}{\Pr(D = d | X)} \right], \\ \delta(d) &= E \left[ \frac{YI\{D = d\}}{\Pr(D = d | M, X)} \left( \frac{\Pr(D = 1 | M, X)}{\Pr(D = 1 | X)} - \frac{1 - \Pr(D = 1 | M, X)}{1 - \Pr(D = 1 | X)} \right) \right], \\ E[Y(d, M(d'))] &= E \left[ \frac{YI\{D = d\}}{\Pr(D = d | M, X)} \frac{\Pr(D = d' | M, X)}{\Pr(D = d' | X)} \right] \text{ for } d, d' \in \{1, 0\}. \end{aligned} \quad (4)$$

$I\{\cdot\}$  is the indicator function which is one if its argument is satisfied and zero otherwise. The expressions for  $\theta(d)$  and  $\delta(d)$  are (by Bayes' theorem) mathematically identical to weighting-based

representations of the direct and indirect effect relying on  $\Pr(D = 1|X)$  and  $\Pr(M = m|D, X)$  (rather than  $\Pr(D = 1|M, X)$ ) suggested in Hong (2010) and Tchetgen Tchetgen and Shpitser (2012), see their ‘strategy 3’. The practical advantage of the approach advocated in this paper is that  $\Pr(D = 1|M, X)$  may be easier to estimate than  $\Pr(M = m|D, X)$  or the respective conditional density of  $M$  in the case of a continuous  $M$ . This particularly relevant when the support of  $M$  is rich (i.e., contains many values) or  $M$  is a vector of several variables.

We note that the identification results in (4) can be generalized in various dimensions. First, replacing  $Y$  everywhere in (4) by an indicator function that  $Y$  is smaller than or equal to a particular value  $a$ , i.e.  $I\{Y \leq a\}$ , allows the evaluation of distributional features and effects. In this case, the expressions for  $\theta(d)$  and  $\delta(d)$  provide the direct and indirect effects on the cumulative distribution function (cdf) evaluated at  $a$  rather than the average effects. Likewise, the expressions for the mean potential outcomes identify the potential cdf’s when  $Y$  is substituted by  $I\{Y \leq a\}$ . Inverting the cdf’s in turn allows identifying quantile treatment effects, given that  $Y$  satisfies particular continuity conditions. See Donald and Hsu (2014) for related results in the context of (total) treatment effect evaluation. Second, the results can be extended to the evaluation of weighted direct and indirect effects as a function of the distribution of  $X$  in an analogous way as suggested in Hirano, Imbens, and Ridder (2003), henceforth HIR, in the context of the ATE. As more thoroughly discussed in Section 4, this permits the identification and (under particular assumptions) root-n-consistent estimation in specific subgroups. As one important special case of weighted IPW, Section 4.3 shows this for the treated population. Finally, the properties derived for our estimators also apply to the estimation of the direct and partial indirect effects discussed in Section 2.3. of Huber (2014) when some covariates in  $X$  are themselves affected by the treatment.

### 3 Estimation and inference

#### 3.1 Nonparametric estimation

As in Huber (2014), the proposed estimators are based on normalized versions of the sample analogs of the IPW-based identification results of expression (4), with weights adding up to unity in either treatment state, see Imbens (2004) and Busso, DiNardo, and McCrary (2014). The normalized estimators of the direct effects under treatment and non-treatment, for instance,

correspond to

$$\begin{aligned}\hat{\theta}(1) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i / \hat{p}(X_i)}{\frac{1}{n} \sum_{i=1}^n D_i / \hat{p}(X_i)} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]}, \\ \hat{\theta}(0) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) / (1 - \hat{p}(X_i))},\end{aligned}\tag{5}$$

and the normalized estimators of the indirect effects under treatment and non-treatment correspond to

$$\begin{aligned}\hat{\delta}(1) &= \frac{\frac{1}{n} \sum_{i=1}^n D_i Y_i / \hat{p}(X_i)}{\frac{1}{n} \sum_{i=1}^n D_i / \hat{p}(X_i)} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}, \\ \hat{\delta}(0) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) / (1 - \hat{p}(X_i))},\end{aligned}\tag{6}$$

$i$  indexes the observations in an i.i.d. sample of size  $n$ .  $\hat{p}(X_i)$  and  $\hat{p}(M_i, X_i)$  denote estimates of the true propensity scores  $\Pr(D = 1|X = X_i)$  and  $\Pr(D = 1|M = M_i, X = X_i)$ , henceforth abbreviated by  $p(X_i)$  and  $p(M_i, X_i)$ , respectively. In contrast to Huber (2014), we estimate the propensity scores nonparametrically by series logit estimation (SLE) based on power series as in HIR. Normalized estimators for the indirect effects, denoted by  $\hat{\delta}(1), \hat{\delta}(0)$ , are obtained in an analogous way. See also Appendix A.1 for the normalized estimators of the potential outcomes.

To illustrate the SLE approach consider, for instance,  $\hat{p}(X_i)$  and suppose that  $X$  contains only continuous variables with dimension  $d_x$ . Let  $\lambda = (\lambda_1, \dots, \lambda_{d_x})' \in \mathbb{Z}_+^{d_x}$  be a  $d_x$ -dimensional vector of non-negative integers where  $\mathbb{Z}_+$  denotes the set of non-negative integers, and define the norm for  $\lambda$  as  $|\lambda| = \sum_{j=1}^{d_x} \lambda_j$ . Furthermore, let  $\{\lambda(k)\}_{k=1}^\infty$  be a sequence including all distinct  $\lambda \in \mathbb{Z}_+^{d_x}$  such that  $|\lambda(k)|$  is non-decreasing in  $k$  and let  $x^\lambda = \prod_{j=1}^{d_x} x_j^{\lambda_j}$ . For any integer  $K_x$ , define  $R^{K_x}(x) = (x^{\lambda(1)}, \dots, x^{\lambda(K_x)})'$  as a vector of power functions. Denote by  $\mathcal{L}(a) = \exp(a)/(1+\exp(a))$  the logistic cumulative distribution function (CDF). The SLE for  $p(x)$  is defined as  $\hat{p}(x) = \mathcal{L}(R^{K_x}(x)' \hat{\pi}_{K_x})$  where

$$\hat{\pi}_{K_x} = \arg \max_{\pi_k} \frac{1}{n} \sum_{i=1}^n \left( D_i \ln \mathcal{L}(R^{K_x}(X_i)' \pi_{K_x}) + (1 - D_i) \ln (1 - \mathcal{L}(R^{K_x}(X_i)' \pi_{K_x})) \right).$$

The asymptotic properties of  $\hat{p}(x)$  are discussed in Appendix A of HIR.  $\hat{p}(m, x)$  is defined in an analogous way.



### 3.2 Asymptotic behaviour

To show root- $n$ -consistency and asymptotic normality of our estimators, we subsequently introduce regularity conditions that are very much related to those in HIR.

**Assumption 4 (distribution of  $(X, M)$ ):**

(i) The distribution of the  $(d_m + d_x)$ -dimensional vector  $(M, X)$  is absolutely continuous with probability density  $f(m, x)$ ; (ii)  $\mathcal{M}$  and  $\mathcal{X}$ , are Cartesian products of compact intervals; (iii)  $f(m, x)$  is twice continuously differentiable, bounded above, and bounded away from 0 on  $\mathcal{M} \times \mathcal{X}$ .

Assumption 4 rules out that  $M$  and/or  $X$  contain any binary or discrete variables, which seems restrictive for empirical applications. We impose this assumption for the sake of ease of discussion and note that our results could be easily extended to cases including discrete covariates and mediators, however, at the cost of more burdensome notation. That is, the sample would need to be stratified on the discrete values and SLE would have to be performed separately for each stratum, see Donald, Hsu, and Lieli (2014).

**Assumption 5 (smoothness of propensity scores):**

(i)  $p(x)$  is continuously differentiable of order  $\bar{p}_x \geq 7d_x$ ; (ii)  $p(m, x)$  is continuously differentiable of order  $\bar{p}_m \geq 7(d_m + d_x)$ .

Assumption 5 is analogous to Assumption 4 of HIR and requires the propensity scores to be sufficiently smooth.

**Assumption 6 (series estimator):**

(i) The SLE of  $p(x)$  uses a power series with  $K_x = N^{\nu_x}$  for some  $d_x/4(\bar{p}_x - d_x) < \nu_x < 1/9$ ,  $K_x^3 n^{-1/2} \rightarrow 0$  and  $K_x^{-(\bar{p}_x + 2d_x)/2d_x} n^{1/4} \rightarrow 0$ ; (ii) The SLE of  $p(m, x)$  uses a power series with  $K_m = N^{\nu_m}$  for some  $(d_m + d_x)/4(\bar{p}_x - d_m - d_x) < \nu_m < 1/9$ ,  $K_m^3 n^{-1/2} \rightarrow 0$  and  $K_m^{-(\bar{p}_m + 2d_m + 2d_x)/(2d_m + 2d_x)} n^{1/4} \rightarrow 0$ .

Assumption 6 restricts the growth rate of the number of approximating functions to be included in the series estimator of the propensity score. Note that our assumption is stronger than Assumption 5 of HIR by requiring that  $K_x^3 n^{-1/2} \rightarrow 0$  and  $K_x^{-(\bar{p}_x + 2d_x)/2d_x} n^{1/4} \rightarrow 0$ , which ensures that  $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$ . An analogous result applies to  $p(m, x)$ . These extra conditions on power series terms are needed because in contrast to HIR, our estimation ap-

proach is based on two propensity scores. When employing mean-value expansions, we require that the second order terms are of order  $o_p(1)$ , for which  $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$  and  $\sup_{m \in \mathcal{M}, x \in \mathcal{X}} |\hat{p}(m, x) - p(m, x)| = o_p(n^{-1/4})$  are sufficient conditions.

We now define several conditional moments related to potential and observed outcomes:  $\rho_{dd'}(x) = E[Y(d, M(d'))|X = x]$  and  $\zeta_d(m, x) = E[Y|M = m, X = x, D = d]$  for  $d, d' \in \{0, 1\}$ . Note that

$$\begin{aligned}\rho_{10}(x) &= E\left[\frac{(1-D)Y}{1-p(X)} \frac{p(M, X)}{(1-p(M, X))} \middle| X = x\right], \\ \rho_{01}(x) &= E\left[\frac{DY}{p(X)} \frac{1-p(M, X)}{p(M, X)} \middle| X = x\right], \\ \zeta_1(m, x) &= E\left[\frac{DY}{p(X, M)} \middle| M = m, X = x\right], \\ \zeta_0(m, x) &= E\left[\frac{(1-D)Y}{1-p(X, M)} \middle| M = m, X = x\right].\end{aligned}$$

Assumption 7 imposes some regularity conditions on these moments and the second moments of the potential outcomes.

**Assumption 7 (moments of  $Y$ ):**

(i)  $E[Y^2(0)] < \infty$  and  $E[Y^2(1)] < \infty$ ; (ii) for  $d, d' \in \{0, 1\}$ ,  $\rho_{dd'}(x)$  are continuously differentiable over  $\mathcal{X}$ ; (iii) for  $d = 0, 1$ ,  $\zeta_d(m, x)$  is continuously differentiable over  $\mathcal{M} \times \mathcal{X}$ .

Under our assumptions, nonparametric IPW estimation of the direct and indirect effects using SLE-based propensity scores is root-n-consistent and asymptotically normal.

**Theorem 1** *Under Assumptions 1 to 7,*

$$\sqrt{n} \begin{pmatrix} \hat{\theta}(1) - \theta(1) \\ \hat{\theta}(0) - \theta(0) \\ \hat{\delta}(1) - \delta(1) \\ \hat{\delta}(0) - \delta(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \mathcal{V}), \quad (7)$$

where  $V$  is a  $4 \times 4$  covariance matrix generated by  $\psi = (\psi_{\theta(1)}, \psi_{\theta(0)}, \psi_{\delta(1)}, \psi_{\delta(0)})'$  with

$$\begin{aligned}
\psi_{\theta(1)}(Y, M, D, X) &= \psi_{11}(Y, M, D, X) - \psi_{01}(Y, M, D, X), \\
\psi_{\theta(0)}(Y, M, D, X) &= \psi_{10}(Y, M, D, X) - \psi_{10}(Y, M, D, X), \\
\psi_{\delta(1)}(Y, M, D, X) &= \psi_{01}(Y, M, D, X) - \psi_{00}(Y, M, D, X), \\
\psi_{\delta(0)}(Y, M, D, X) &= \psi_{10}(Y, M, D, X) - \psi_{00}(Y, M, D, X), \\
\psi_{11}(Y, M, D, X) &= \frac{DY}{p(X)} - \frac{\rho_{11}(X)}{p(X)}(D - p(X)) - \mu_{11}, \\
\psi_{00}(Y, M, D, X) &= \frac{(1-D)Y}{1-p(X)} + \frac{\rho_{00}(X)}{1-p(X)}(D - p(X)) - \mu_{00}, \\
\psi_{10}(Y, M, D, X) &= \frac{DY}{p(M, X)} \frac{1-p(M, X)}{1-p(X)} + \frac{\rho_{10}(X)}{(1-p(X))}(D - p(X)) \\
&\quad - \frac{\zeta_1(M, X)}{p(M, X)(1-p(X))}(D - p(M, X)) - \mu_{10}, \\
\psi_{01}(Y, M, D, X) &= \frac{(1-D)Y}{1-p(M, X)} \frac{p(M, X)}{p(X)} - \frac{\rho_{01}(X)}{p(X)}(D - p(X)) \\
&\quad + \frac{\zeta_0(M, X)}{(1-p(M, X))p(X)}(D - p(M, X)) - \mu_{01},
\end{aligned}$$

and  $\mu_{dd'} = E[Y(d, M(d'))]$  for  $d, d' \in \{0, 1\}$ .

We note that the expressions in Theorem 1 imply that our estimators attain the semiparametric efficiency bounds for mediation analysis derived in Tchetgen Tchetgen and Shpitser (2012). Finally, a relevant question for the practical implementation of the estimator is how to choose orders  $K_x$  and  $K_m$  for propensity score estimation. In the simulations presented in Section 5, we consider cross-validation for picking either parameter, as well as overfitting by one order higher than suggested by cross-validation. Even though cross-validation w.r.t. the the functional form of the propensity score does in general not provide the optimal order for the estimation of direct and indirect effects in a given sample, the simulation results suggest that this approach can yield satisfactory results in practice.

### 3.3 Inference

Inference based on the asymptotic results in Theorem 1 requires a consistent estimator of the asymptotic covariance matrix, denoted by  $\mathcal{V}$ . We first propose uniformly consistent estimators for  $\rho_{dd'}(x)$  and  $\zeta_d(m, x)$  with  $d, d' \in \{0, 1\}$ . Let  $R^{K_x}(x)$  and  $R^{K_m}(m, x)$  be the column vectors of

power functions used for the estimation of the propensity score functions and define

$$\begin{aligned}
\hat{\rho}_{11}(x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(X_i)} R^{K_x}(X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_x}(X_i) R^{K_x}(X_i)' \right)^{-1} R^{K_x}(x), \\
\hat{\rho}_{00}(x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(X_i)} R^{K_x}(X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_x}(X_i) R^{K_x}(X_i)' \right)^{-1} R^{K_x}(x), \\
\hat{\rho}_{10}(x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{D_i Y}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} R^{K_x}(X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_x}(X_i) R^{K_x}(X_i)' \right)^{-1} R^{K_x}(x), \\
\hat{\rho}_{01}(x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} R^{K_x}(X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_x}(X_i) R^{K_x}(X_i)' \right)^{-1} R^{K_x}(x), \\
\hat{\zeta}_1(m, x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(M_i, X_i)} R^{K_m}(M_i, X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_m}(M_i, X_i) R^{K_m}(M_i, X_i)' \right)^{-1} R^{K_m}(m, x), \\
\hat{\zeta}_0(m, x) &= \left( \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(M_i, X_i)} R^{K_m}(M_i, X_i)' \right) \left( \frac{1}{n} \sum_{i=1}^n R^{K_m}(M_i, X_i) R^{K_m}(M_i, X_i)' \right)^{-1} R^{K_m}(m, x).
\end{aligned}$$

This permits defining the estimated influence functions as

$$\begin{aligned}
\hat{\psi}_{\theta(1)}(Y, M, D, X) &= \hat{\psi}_{11}(Y, M, D, X) - \hat{\psi}_{01}(Y, M, D, X), \\
\hat{\psi}_{\theta(0)}(Y, M, D, X) &= \hat{\psi}_{10}(Y, M, D, X) - \hat{\psi}_{00}(Y, M, D, X), \\
\hat{\psi}_{\delta(1)}(Y, M, D, X) &= \hat{\psi}_{01}(Y, M, D, X) - \hat{\psi}_{00}(Y, M, D, X), \\
\hat{\psi}_{\delta(0)}(Y, M, D, X) &= \hat{\psi}_{10}(Y, M, D, X) - \hat{\psi}_{00}(Y, M, D, X), \\
\hat{\psi}_{11}(Y, M, D, X) &= \frac{DY}{\hat{p}(X)} - \frac{\hat{\rho}_{11}(X)}{\hat{p}(X)} (D - \hat{p}(X)) - \hat{\mu}_{11}, \\
\hat{\psi}_{00}(Y, M, D, X) &= \frac{(1-D)Y}{1-\hat{p}(X)} + \frac{\hat{\rho}_{00}(X)}{1-\hat{p}(X)} (D - \hat{p}(X)) - \hat{\mu}_{00}, \\
\hat{\psi}_{10}(Y, M, D, X) &= \frac{DY}{\hat{p}(M, X)} \frac{1-\hat{p}(M, X)}{1-\hat{p}(X)} + \frac{\hat{\rho}_{10}(X)}{(1-\hat{p}(X))} (D - \hat{p}(X)) \\
&\quad - \frac{\hat{\zeta}_1(M, X)}{\hat{p}(M, X)(1-\hat{p}(X))} (D - \hat{p}(M, X)) - \hat{\mu}_{10}, \\
\hat{\psi}_{01}(Y, M, D, X) &= \frac{(1-D)Y}{1-\hat{p}(M, X)} \frac{\hat{p}(M, X)}{\hat{p}(X)} - \frac{\hat{\rho}_{01}(X)}{\hat{p}(X)} (D - \hat{p}(X)) \\
&\quad + \frac{\hat{\zeta}_0(M, X)}{(1-\hat{p}(M, X))\hat{p}(X)} (D - \hat{p}(M, X)) - \hat{\mu}_{01},
\end{aligned}$$

where

$$\begin{aligned}\hat{\mu}_{11} &= \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\hat{p}(X_i)}, & \hat{\mu}_{00} &= \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{10} &= \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{01} &= \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)}.\end{aligned}$$

Furthermore, define  $\hat{\psi}_i = (\hat{\psi}_{\theta(1),i}, \hat{\psi}_{\theta(0),i}, \hat{\psi}_{\delta(1),i}, \hat{\psi}_{\delta(0),i})'$  with

$$\begin{aligned}\hat{\psi}_{\theta(1),i} &= \hat{\psi}_{11}(Y_i, M_i, D_i, X_i) - \hat{\psi}_{01}(Y_i, M_i, D_i, X_i), \\ \hat{\psi}_{\theta(0),i} &= \hat{\psi}_{10}(Y_i, M_i, D_i, X_i) - \hat{\psi}_{10}(Y_i, M_i, D_i, X_i), \\ \hat{\psi}_{\delta(1),i} &= \hat{\psi}_{01}(Y_i, M_i, D_i, X_i) - \hat{\psi}_{00}(Y_i, M_i, D_i, X_i), \\ \hat{\psi}_{\delta(0),i} &= \hat{\psi}_{10}(Y_i, M_i, D_i, X_i) - \hat{\psi}_{00}(Y_i, M_i, D_i, X_i).\end{aligned}$$

Finally, let  $\hat{\mathcal{V}} = n^{-1} \sum_{i=1}^n \hat{\psi}_i \hat{\psi}_i'$ . By the same arguments as in HIR, one can show that  $\hat{\mathcal{V}} \xrightarrow{P} \mathcal{V}$ . The details are therefore omitted. We also note that given the asymptotic normality of our estimators, the bootstrap is a consistent inference method, too, and is used in our application discussed further below.

## 4 Weighted direct and indirect effects

### 4.1 Identification of weighted effects

In this section, we discuss the identification, estimation, and asymptotic results for weighted effects. Let to this end  $g(X)$  denote a weighting function that depends on  $X$  or subsets thereof and satisfies  $|g(X)| < \infty$  and  $E[g(X)] > 0$ . Weighted direct and indirect effects as well as mean potential outcomes (denoted as  $\theta_g(d)$ ,  $\delta_g(d)$ ,  $E_g[Y(d, M(d'))]$ ) are identified by including

$g(X)/E[g(X)]$  in the respective expectation operators presented in (4):

$$\begin{aligned}
\theta_g(d) &= E \left[ \frac{g(X)}{E[g(X)]} \left( \frac{YD}{\Pr(D=1|M, X)} - \frac{Y(1-D)}{1 - \Pr(D=1|M, X)} \right) \frac{\Pr(D=d|M, X)}{\Pr(D=d|X)} \right], \\
\delta_g(d) &= E \left[ \frac{g(X)}{E[g(X)]} \frac{YI\{D=d\}}{\Pr(D=d|M, X)} \left( \frac{\Pr(D=1|M, X)}{\Pr(D=1|X)} - \frac{1 - \Pr(D=1|M, X)}{1 - \Pr(D=1|X)} \right) \right]. \\
E_g[Y(d, M(d'))] &= E \left[ \frac{g(X)}{E[g(X)]} \frac{YI\{D=d\}}{\Pr(D=d|M, X)} \frac{\Pr(D=d'|M, X)}{\Pr(D=d'|X)} \right] \text{ for } d, d' \in \{1, 0\}. \quad (8)
\end{aligned}$$

This allows identifying the effects for specific subgroups of interest. For instance, setting  $g(X) = p(X)$  with  $E[g(X)] = \Pr(D=1)$  yields the direct and indirect effects as well as the potential outcomes among the treated. To see this, consider  $E_g[Y(d, M(d'))] = E[Y(d, M(d'))|D=1]$  and note that (when  $M, X$  are continuous),

$$\begin{aligned}
& E \left[ \frac{p(X)}{E[p(X)]} \frac{YI\{D=d\}}{\Pr(D=d|M, X)} \frac{\Pr(D=d'|M, X)}{\Pr(D=d'|X)} \right] \quad (9) \\
&= E \left[ \frac{p(X)}{\Pr(D=1)} E \left[ E \left[ \frac{YI\{D=d\}}{\Pr(D=d|M, X)} \middle| M=m, X=x \right] \frac{\Pr(D=d'|M, X)}{\Pr(D=d'|X)} \middle| X=x \right] \right] \\
&= \int \frac{p(X)}{\Pr(D=1)} \int E[Y|D=d, M=m, X=x] \frac{\Pr(D=d'|M, X)}{\Pr(D=d'|X)} dF_{M|X=x}(m) dF_X(x) \\
&= \int \frac{p(X)}{\Pr(D=1)} \int E[Y(d, m)|D=d, M=m, X=x] dF_{M|D=d, X=x}(m) dF_X(x) \\
&= \int \frac{p(X)}{\Pr(D=1)} \int E[Y(d, m)|M(d')=m, X=x] dF_{M(d')|X=x}(m) dF_X(x) \\
&= \int E[Y(d, M(d'))|X=x] \frac{p(X)}{\Pr(D=1)} dF_X(x) \\
&= \int E[Y(d, M(d'))|X=x] dF_{X|D=1}(x) \\
&= E[Y(d, M(d'))|D=1].
\end{aligned}$$

The first equation follows from the law of iterated expectations, the second from basic probability theory and replacing expectations by integrals, the third and sixth from Bayes' theorem, the fourth from Assumptions 1 and 2, and the fifth and seventh from integration. Analogously, the parameters for the nontreated are obtained by setting  $g(X) = 1 - p(X)$ . Furthermore, this approach can be used to assess effect heterogeneity w.r.t.  $X$ , e.g. by defining  $g(X)$  as an indicator function that  $X$  takes particular (ranges of) values.

## 4.2 Estimation and asymptotics under known weighting functions

Based on (8), we propose the following estimators for a known weighting function  $g(X)$ :

$$\begin{aligned}
\hat{\theta}_g(1) &= \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i D_i / \hat{p}(X_i)}{\frac{1}{n} \sum_{i=1}^n g(X_i) D_i / \hat{p}(X_i)} - \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]}{\frac{1}{n} \sum_{i=1}^n g(X_i) (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]} \\
\hat{\theta}_g(0) &= \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n g(X_i) D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]} - \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i (1 - D_i) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) / (1 - \hat{p}(X_i))}, \\
\hat{\delta}_g(1) &= \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) D_i Y_i / \hat{p}(X_i)}{\frac{1}{n} \sum_{i=1}^n g(X_i) D_i / \hat{p}(X_i)} - \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n g(X_i) D_i (1 - \hat{p}(M_i, X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}, \quad (10) \\
\hat{\delta}_g(0) &= \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]}{\frac{1}{n} \sum_{i=1}^n g(X_i) (1 - D_i) \hat{p}(M_i, X_i) / [(1 - \hat{p}(M_i, X_i)) \hat{p}(X_i)]} - \frac{\frac{1}{n} \sum_{i=1}^n g(X_i) Y_i (1 - D_i) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) / (1 - \hat{p}(X_i))},
\end{aligned}$$

Under our previously discussed assumptions, nonparametric IPW estimation of the direct and indirect effects using SLE-based propensity scores are root- $n$ -consistent and asymptotically normal.

**Theorem 2** *Suppose that  $|g(x)| < M < \infty$  and  $E[g(X)] > 0$ . Under Assumptions 1 to 7,*

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_g(1) - \theta_g(1) \\ \hat{\theta}_g(0) - \theta_g(0) \\ \hat{\delta}_g(1) - \delta_g(1) \\ \hat{\delta}_g(0) - \delta_g(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \mathcal{V}_g) \quad (11)$$

where  $V_g$  is a  $4 \times 4$  covariance matrix generated by  $\psi_g = (\psi_{\theta_g(1)}, \psi_{\theta_g(0)}, \psi_{\delta_g(1)}, \psi_{\delta_g(0)})'$  with

$$\begin{aligned}
\psi_{\theta_g(1)}(Y, M, D, X) &= \psi_{11,g}(Y, M, D, X) - \psi_{01,g}(Y, M, D, X) \\
\psi_{\theta_g(0)}(Y, M, D, X) &= \psi_{10,g}(Y, M, D, X) - \psi_{10,g}(Y, M, D, X) \\
\psi_{\delta_g(1)}(Y, M, D, X) &= \psi_{01,g}(Y, M, D, X) - \psi_{00,g}(Y, M, D, X) \\
\psi_{\delta_g(0)}(Y, M, D, X) &= \psi_{10,g}(Y, M, D, X) - \psi_{00,g}(Y, M, D, X) \\
\psi_{11,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]} \left( \frac{DY}{p(X)} - \frac{\rho_{11}(X)}{p(X)}(D - p(X)) - \mu_{11,g} \right), \\
\psi_{00,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]} \left( \frac{(1-D)Y}{1-p(X)} + \frac{\rho_{00}(X)}{1-p(X)}(D - p(X)) - \mu_{00,g} \right), \\
\psi_{10,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]} \left( \frac{DY}{p(M, X)} \frac{1-p(M, X)}{1-p(X)} + \frac{\rho_{10}(X)}{(1-p(X))} (D - p(X)) \right. \\
&\quad \left. - \frac{\zeta_1(M, X)}{P(M, X)(1-P(X))} (D - p(M, X)) - \mu_{10,g} \right), \\
\psi_{01,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]} \left( \frac{(1-D)Y}{1-p(M, X)} \frac{p(M, X)}{p(X)} - \frac{\rho_{01}(X)}{p(X)} (D - p(X)) \right. \\
&\quad \left. + \frac{\zeta_0(M, X)}{(1-p(M, X))p(X)} (D - p(M, X)) - \mu_{01,g} \right),
\end{aligned}$$

where  $\mu_{dd',g} = E[g(X)Y(d, M(d'))]/E[g(X)]$  for  $d, d' \in \{0, 1\}$ .

Inference for weighted effects can be performed in a similar way as outlined in Section 3.3.

### 4.3 Effects on the treated with estimated propensity scores

In this section, we discuss the estimation and asymptotic results for the subgroup of treated. Note that if the propensity score  $p(X)$  was known, the results of Section 4.2 with  $g(X) = p(X)$  would immediately apply. However, practically more relevant is the case that  $p(X)$  is unknown and needs to be estimated, which is considered in this section. We denote by  $\hat{\theta}_t(d)$  and  $\hat{\delta}_t(d)$  estimates of the direct and indirect effects among the treated,  $\theta_t(d) = E[Y(1, M(d)) - Y(0, M(d)) | D = 1]$  and  $\delta_t(d) = E[Y(d, M(1)) - Y(d, M(0)) | D = 1]$ , respectively.

The normalized sample analogs of the effects in (8) with an unknown weighting function



$g(X) = p(X)$  correspond to:

$$\begin{aligned}
\hat{\theta}_t(1) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i}{\frac{1}{n} \sum_{i=1}^n D_i} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) \hat{p}(M_i, X_i) / (1 - \hat{p}(M_i, X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) \hat{p}(M_i, X_i) / (1 - \hat{p}(M_i, X_i))}, \\
\hat{\theta}_t(0) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i (1 - \hat{p}(M_i, X_i)) (\hat{p}(X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n D_i (1 - \hat{p}(M_i, X_i)) (\hat{p}(X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]} \\
&\quad - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) (\hat{p}(X_i)) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) (\hat{p}(X_i)) / (1 - \hat{p}(X_i))}, \\
\hat{\delta}_t(1) &= \frac{\frac{1}{n} \sum_{i=1}^n D_i Y_i}{\frac{1}{n} \sum_{i=1}^n D_i} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i D_i (1 - \hat{p}(M_i, X_i)) (\hat{p}(X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}{\frac{1}{n} \sum_{i=1}^n D_i (1 - \hat{p}(M_i, X_i)) (\hat{p}(X_i)) / [\hat{p}(M_i, X_i) (1 - \hat{p}(X_i))]}, \\
\hat{\delta}_t(0) &= \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) \hat{p}(M_i, X_i) / (1 - \hat{p}(M_i, X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) \hat{p}(M_i, X_i) / (1 - \hat{p}(M_i, X_i))} - \frac{\frac{1}{n} \sum_{i=1}^n Y_i (1 - D_i) (\hat{p}(X_i)) / (1 - \hat{p}(X_i))}{\frac{1}{n} \sum_{i=1}^n (1 - D_i) (\hat{p}(X_i)) / (1 - \hat{p}(X_i))},
\end{aligned} \tag{12}$$

Under the same assumptions as before, nonparametric IPW estimation of the direct and indirect effects among the treated based on SLE-based propensity scores is root-n-consistent and asymptotically normal.

**Theorem 3** *Under Assumptions 1 to 7,*

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_t(1) - \theta_t(1) \\ \hat{\theta}_t(0) - \theta_t(0) \\ \hat{\delta}_t(1) - \delta_t(1) \\ \hat{\delta}_t(0) - \delta_t(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \mathcal{V}_t) \tag{13}$$

where  $V_t$  is a  $4 \times 4$  covariance matrix generated by  $\psi_t = (\psi_{\theta_t(1)}, \psi_{\theta_t(0)}, \psi_{\delta_t(1)}, \psi_{\delta_t(0)})'$  with

$$\begin{aligned}
\psi_{\theta_t(1)}(Y, M, D, X) &= \psi_{11,t}(Y, M, D, X) - \psi_{01,t}(Y, M, D, X) \\
\psi_{\theta_t(0)}(Y, M, D, X) &= \psi_{10,t}(Y, M, D, X) - \psi_{00,t}(Y, M, D, X) \\
\psi_{\delta_t(1)}(Y, M, D, X) &= \psi_{01,t}(Y, M, D, X) - \psi_{00,t}(Y, M, D, X) \\
\psi_{\delta_t(0)}(Y, M, D, X) &= \psi_{10,t}(Y, M, D, X) - \psi_{00,t}(Y, M, D, X) \\
\psi_{11,t}(Y, M, D, X) &= \frac{1}{E[p(X)]} \left( D(Y - \mu_{11,t}) \right), \\
\psi_{00,t}(Y, M, D, X) &= \frac{1}{E[p(X)]} \left( \frac{p(X)(1-D)(Y - \rho_{00}(X))}{1-p(X)} + (\rho_{00}(X) - \mu_{00,t})D \right), \\
\psi_{10,t}(Y, M, D, X) &= \frac{1}{E[p(X)]} \left( \frac{p(X)DY}{p(M, X)} \frac{1-p(M, X)}{1-p(X)} + \frac{\rho_{10}(X)}{1-p(X)}(D-p(X)) \right. \\
&\quad \left. - \frac{p(X)\zeta_1(M, X)}{p(M, X)(1-p(X))}(D-p(M, X)) - \mu_{10,t}D \right), \\
\psi_{01,t}(Y, M, D, X) &= \frac{1}{E[p(X)]} \left( \frac{(1-D)Y}{1-p(M, X)}p(M, X) + \frac{\zeta_0(M, X)}{1-p(M, X)}(D-p(M, X)) - \mu_{01,t}D \right),
\end{aligned}$$

where  $\mu_{dd',t} = E[p(X)Y(d, M(d'))]/E[p(X)]$  for  $d, d' \in \{0, 1\}$ .

## 5 Simulations

This section presents a brief simulation study in which we investigate the finite sample performance of nonparametric IPW based on (5) and SLE of the propensity scores, as well as of alternative estimators by considering the following data generating process:

$$\begin{aligned}
D &= I\{\beta(X_1^2 + X_2) + \epsilon_D > 0\}, \quad M = I\{\beta(D + X_1^2 + X_2) + \epsilon_M > 0\}, \\
Y &= D + M + \beta[(1+D)(X_1^2 + X_2) + DM(1 + X_1^2 + X_2)] + \epsilon_Y,
\end{aligned} \tag{14}$$

with  $X_1, \epsilon_D, \epsilon_M, \epsilon_Y \sim \mathcal{N}(0, 1)$ ,  $X_2 \sim \text{binomial}(0.5)$ , independently of each other.

$X_1$  and  $X_2$  are observed covariates and follow standard normal and binomial distributions, respectively. We note that  $X_1$  enters the equation of the continuous outcome  $Y$  and the index functions of the binary treatment and mediator variables  $D$  and  $M$  both linearly and quadratically.  $\epsilon_D, \epsilon_M, \epsilon_Y$  are random and standard normally distributed unobservables. The parameter  $\beta$  gauges the degree of confounding, i.e. how strongly the covariates jointly affect  $D, M$ , and  $Y$ .  $\beta$  also determines the level of effect heterogeneity across values of the mediator

and the covariates rooted in the interaction terms between  $D$ ,  $M$ ,  $X_1$ , and  $X_2$  in the outcome equation. In our simulations with 1000 replications, we set  $\beta = 0.1, 1.5, 0.2$ , implying low and somewhat stronger confounding and effect heterogeneity, and consider two sample sizes of  $n = 1000, 4000$ .<sup>4</sup>

We investigate the performance of the following estimators: (i) Nonparametric IPW using SLE-based propensity scores as outlined in Section 3, where the orders  $K_x$  and  $K_m$  for the series approximations are chosen by 10-fold cross-validation for either propensity score given  $X_1, X_2$  and  $X_1, X_2, M$ , respectively. For implementation, we make use of some functions provided in the ‘LARF’ package of An and Wang (2016) for the statistical software R. On top of IPW based on cross-validated propensity scores (ipw cv), we also consider an overfitted version (ipw ofit) where  $K_x$  and  $K_m$  are one order higher than suggested by cross-validation. (ii) Semiparametric IPW based on parametric plug-in estimators as in Huber (2014) (semi ipw), using probit models for the treatment propensity scores. Note that all IPW estimators of (i) and (ii) apply a trimming rule that discards observations with propensity scores smaller than 0.02 or larger than 0.98 to prevent exploding weights due to small denominators.<sup>5</sup>

(iii) g-computation (g.comp) as suggested by Robins (1986) and used in the context of direct and indirect effects for instance in Zheng and van der Laan (2012). It is based on maximum likelihood estimation of the mediator and outcome models as logit and linear functions of  $D, X_1, X_2$  and  $D, M, DM, X_1, X_2$ , respectively. It therefore omits the squared terms of  $X_1$  as well as interactions with covariates in the outcome equation. (iv) Multiply robust estimation (mr) as suggested in Tchetgen Tchetgen and Shpitser (2012) based on the sample analog of the efficient influence or score function, which relies on estimates of the treatment, mediator, and outcome models. The latter two follow the same specifications as in g-computation, while  $D$  is modelled as a logit function of  $X_1, X_2$ . (v) Simulation-based estimation (sim) as proposed by Tingley, Yamamoto, Hirose, Imai, and Keele (2014) and implemented in the ‘mediation’ package for R. It is based on estimating the mediator and outcome models using the same (logit and linear) models as in g-computation and simulating potential mediators and outcomes to compute

---

<sup>4</sup>For  $\beta = 0.1$ , the averages of  $D$  and  $M$  are 0.56 and 0.58, respectively, in our simulations. For  $\beta = 0.2$ , the averages of  $D$  and  $M$  are 0.65 and 0.61, respectively.

<sup>5</sup>Only few observations in our simulations need to be trimmed. For ( $\beta = 0.1, n = 1000$ ), on average 0.25 overfitted SLE-based propensity scores including both the covariates and the mediator lie outside the  $[0.02, 0.98]$  interval. This is also the case for on average 0.04 cross-validated SLE-based propensity scores. For ( $\beta = 0.2, n = 1000$ ), the respective numbers are 2.80 and 1.39 for the overfitted and cross-validated SLE-based propensity scores, respectively. None of the probit-based propensity scores are trimmed in any simulation design.

Table 1: Simulation results for  $\beta = 0.1, 1.5, 0.2$  and  $n = 1000, 4000$

$n = 1000$	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	bias	std.dev	RMSE	bias	std.dev	RMSE	bias	std.dev	RMSE	bias	std.dev	RMSE
$\beta = 0.1$												
ipw cv	0.02	0.07	0.07	0.04	0.09	0.10	-0.03	0.07	0.08	-0.01	0.04	0.04
ipw ofit	0.01	0.07	0.07	0.01	0.07	0.07	-0.01	0.04	0.04	-0.01	0.04	0.04
semi ipw	0.06	0.07	0.09	0.05	0.07	0.08	0.02	0.05	0.05	0.00	0.03	0.03
g.comp	0.04	0.07	0.08	0.04	0.07	0.08	0.02	0.05	0.05	0.02	0.05	0.05
mr	0.05	0.07	0.08	0.05	0.07	0.08	0.02	0.05	0.05	0.02	0.05	0.05
sim	0.06	0.07	0.09	0.05	0.07	0.08	0.02	0.05	0.05	0.00	0.04	0.04
$\beta = 0.15$												
ipw cv	0.02	0.08	0.08	0.02	0.09	0.09	-0.02	0.07	0.07	-0.02	0.06	0.06
ipw ofit	0.02	0.08	0.08	0.00	0.07	0.07	-0.00	0.05	0.05	-0.02	0.06	0.06
semi ipw	0.13	0.07	0.15	0.09	0.07	0.11	0.04	0.05	0.07	0.00	0.03	0.03
g.comp	0.08	0.07	0.11	0.08	0.07	0.11	0.04	0.05	0.07	0.04	0.05	0.07
mr	0.09	0.07	0.11	0.09	0.07	0.11	0.04	0.05	0.07	0.04	0.05	0.07
sim	0.13	0.07	0.14	0.09	0.07	0.12	0.03	0.05	0.06	-0.01	0.03	0.04
$\beta = 0.2$												
ipw cv	0.03	0.07	0.08	0.01	0.08	0.08	-0.01	0.06	0.06	-0.03	0.04	0.05
ipw ofit	0.02	0.07	0.08	-0.00	0.08	0.08	-0.00	0.05	0.05	-0.03	0.04	0.05
semi ipw	0.21	0.08	0.22	0.13	0.08	0.15	0.08	0.06	0.10	0.00	0.04	0.04
g.comp	0.12	0.08	0.14	0.12	0.08	0.14	0.08	0.06	0.10	0.08	0.06	0.10
mr	0.13	0.08	0.15	0.13	0.08	0.15	0.08	0.06	0.10	0.08	0.06	0.10
sim	0.20	0.08	0.21	0.13	0.08	0.15	0.04	0.06	0.08	-0.02	0.03	0.04
$n = 4000$												
$\beta = 0.1$												
ipw cv	0.01	0.03	0.03	0.00	0.03	0.03	0.00	0.02	0.02	-0.01	0.02	0.02
ipw ofit	0.01	0.03	0.03	0.00	0.03	0.03	0.00	0.02	0.02	-0.01	0.02	0.02
semi ipw	0.06	0.03	0.07	0.05	0.03	0.06	0.02	0.02	0.03	0.00	0.02	0.02
g.comp	0.04	0.03	0.06	0.04	0.03	0.06	0.02	0.02	0.03	0.02	0.02	0.03
mr	0.05	0.03	0.06	0.05	0.03	0.06	0.02	0.02	0.03	0.02	0.02	0.03
sim	0.06	0.03	0.07	0.05	0.03	0.06	-0.01	0.02	0.02	-0.02	0.02	0.03
$\beta = 0.15$												
ipw cv	0.02	0.03	0.04	0.00	0.03	0.03	0.00	0.02	0.02	-0.02	0.02	0.02
ipw ofit	0.02	0.03	0.04	0.00	0.03	0.03	0.00	0.02	0.02	-0.02	0.02	0.02
semi ipw	0.13	0.04	0.13	0.09	0.03	0.09	0.05	0.02	0.05	0.01	0.02	0.02
g.comp	0.08	0.03	0.09	0.08	0.03	0.09	0.05	0.02	0.05	0.05	0.02	0.05
mr	0.09	0.03	0.09	0.09	0.03	0.09	0.05	0.02	0.05	0.05	0.02	0.05
sim	0.12	0.04	0.13	0.09	0.03	0.10	0.02	0.02	0.03	-0.02	0.02	0.02
$\beta = 0.2$												
ipw cv	0.03	0.04	0.05	0.00	0.04	0.04	-0.00	0.02	0.02	-0.03	0.02	0.04
ipw ofit	0.03	0.04	0.05	-0.00	0.04	0.04	-0.00	0.03	0.03	-0.03	0.02	0.03
semi ipw	0.21	0.04	0.21	0.13	0.04	0.13	0.09	0.03	0.09	0.00	0.02	0.02
g.comp	0.12	0.04	0.13	0.12	0.04	0.13	0.09	0.03	0.09	0.09	0.03	0.09
mr	0.13	0.04	0.13	0.13	0.04	0.13	0.09	0.03	0.09	0.09	0.03	0.09
sim	0.20	0.04	0.20	0.13	0.04	0.14	0.05	0.03	0.06	-0.02	0.02	0.02

Note: ‘bias’, ‘std.dev’, and ‘RMSE’ denote the bias, standard deviation, and root mean squared error. ‘ipw cv’, ‘ipw ofit’, ‘semi ipw’, ‘g.comp’, ‘mr’ and ‘sim’ denote IPW using SLE based on cross-validation, IPW using SLE based on overfitting, semiparametric IPW using probit, g-computation, multiply robust estimation using the efficient influence function, and simulation-based estimation, respectively.

direct and indirect effects.

Table 1 reports the bias, standard deviation (std.dev), and root mean squared error (RMSE) of the estimators for various combinations of  $\beta$  and  $n$ . As a general pattern, nonparametric IPW becomes relatively more competitive when compared to the other estimators as  $n$  and/or  $\beta$  increase. In any simulation design with  $n = 4000$ , our procedure with both cross-validated or overfitted propensity scores by and large dominates the other estimators in terms of RMSE. This is driven by the lower bias of nonparametric IPW due to not imposing any parametric functional form assumptions on the propensity score (in particular when  $\beta$  is large), while all methods are quite comparable in terms of standard deviations. We also see that for  $n = 1000$ , the overfitted version of our method somewhat outperforms estimation based on cross-validation. Under the larger sample size, however, both methods appear to work equally well in terms of bias, standard deviation, and RMSE.

## 6 Application

We apply our methods to experimental data from Chinkhumba, Godlonton, and Thornton (2014) who aim measuring the demand for adult medical male circumcision among 1,634 uncircumcised men in urban Malawi as a function of randomized subsidies for circumcision and comprehensive information on circumcision and HIV. In our mediation analysis, we focus on the information campaign only, which was randomized independently of the financial subsidies. Men receiving comprehensive information were informed that circumcision is partially protective against HIV transmission (based on other empirical studies) at the baseline survey in 2010, while those who did not receive it were only told about the (circumcision) services of the experimenters' partner clinic. We are interested in the effect of information ( $D$ ) on a binary outcome taking the value one if a male has already been circumcised or claimed to be willing to ever get circumcised at the follow up survey ( $Y$ ) in 2011, roughly one year after the baseline survey and treatment assignment.<sup>6</sup> We aim at disentangling the impact of information into a direct effect an indirect component operating through (a change in) the risk assessment of HIV with and without circumcision. To this end, the mediator is defined to be a dummy for whether uncircumcised males are considered

---

<sup>6</sup>To be concise,  $Y$  is defined as one if either the interviewee stated to be circumcised or was willing to ever get circumcised, or the administrative records of the partner clinic show that he actually got circumcised, see the discussion in Chinkhumba, Godlonton, and Thornton (2014).

Table 2: Descriptives

	$D = 1$		$D = 0$		$M = 1$		$M = 0$	
	mean	std.dev	mean	std.dev	mean	std.dev	mean	std.dev
age (in years; discrete)	26.72	5.33	26.43	6.11	26.48	5.80	27.19	5.10
12 years of education (binary)	0.40	0.49	0.42	0.49	0.41	0.49	0.38	0.49
13+ years of education (binary)	0.18	0.38	0.21	0.40	0.19	0.39	0.21	0.41
ever had sex (binary)	0.87	0.34	0.88	0.32	0.88	0.33	0.85	0.35
used condom at last sex (binary)	0.38	0.48	0.42	0.49	0.41	0.49	0.31	0.46
expenditures (in 1000 MKW; continuous)	21.72	32.47	21.98	25.61	21.04	23.59	26.65	51.91
has a working tv / stereo system (binary)	0.78	0.42	0.78	0.41	0.79	0.41	0.76	0.43
has a working car (binary)	0.13	0.34	0.15	0.36	0.14	0.35	0.13	0.34
Y: got or would get circumcised (binary)	0.76	0.43	0.76	0.43	0.80	0.40	0.48	0.50

Note: Sample consists of 1,147 men without missing information in covariates, mediators, and outcomes (note that there are no missing values in the randomly assigned treatment). ‘mean’ and ‘std.dev’ denotes the mean and standard deviation, respectively. \*: ‘believe about HIV risk’ is 0 if individual believes circumcised men to have higher HIV risk than uncircumcised men, 1 if the risk is believed to be the same for both groups, and 2 if uncircumcised men are believed to bear a higher HIV risk.

to be more prone to HIV risk than circumcised ones in the follow up survey ( $M$ ). We therefore aim at evaluating whether it is HIV risk assessment or other factors that constitute important causal mechanisms of the information treatment.

We control for the following baseline covariates ( $X$ ) in our estimation, in particular to account for the endogeneity of the mediator, which is in contrast to the treatment not randomly assigned: age, education (measured in categories of up to 11 years of education, 12 years, or 13 years and more), a dummy for having ever had sex, a dummy for condom use at last sex as a proxy for sexual (risk) behavior, monthly expenditures in Malawian Kwacha (MKW), and dummies for having a working TV and/or stereo system as well as a working car as wealth proxies. We confine our sample to those 1,147 men without missing information in the covariates, the mediator, and the outcome (while there are no missing values in the randomly assigned treatment). Table 2 provides descriptive statistics (means and standard deviations) for the covariates and the outcome across the various treatment and mediator states. We refer to Chinkhumba, Godlonton, and Thornton (2014) for more details about the survey design, the variables, and attrition patterns.

Table 3 reports the direct and indirect effects based on the same estimators as considered in the simulations. In the case of IPW, we again discard observations with extreme propensity scores using the 0.02 trimming rule. No observations are affected by trimming when estimating the propensity score by probit or based on SLE with cross-validation, while 2 subjects are discarded

Table 3: Application

	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	est	se	pval	est	se	pval	est	se	pval	est	se	pval
ipw cv	-0.00	0.02	0.89	-0.01	0.02	0.74	0.01	0.01	0.08	0.01	0.01	0.09
ipw ofit	-0.00	0.03	1.00	-0.01	0.03	0.72	0.01	0.01	0.19	0.01	0.01	0.57
semi ipw	-0.00	0.02	0.89	-0.01	0.02	0.74	0.01	0.01	0.08	0.01	0.01	0.09
g.comp	-0.01	0.02	0.73	-0.01	0.02	0.73	0.01	0.01	0.08	0.01	0.01	0.08
mr	-0.01	0.02	0.73	-0.01	0.02	0.73	0.01	0.01	0.08	0.01	0.01	0.08
sim	-0.00		0.86	-0.01		0.70	0.01		0.12	0.01		0.11

Note: ‘est’, ‘se’, and ‘pval’ denote the effect estimate, the standard error, and the p-value, respectively. ‘ipw cv’, ‘ipw ofit’, ‘semi ipw’, ‘g.comp’, ‘mr’, and ‘sim’ denote IPW using SLE based on cross-validation, IPW using SLE based on overfitting, semiparametric IPW using probit, g-computation, multiply robust estimation using the efficient influence function, and simulation-based estimation, respectively. Standard errors and/or p-values are based on 1999 bootstrap replications or simulations (in the case of ‘sim’).

when using an overfitted propensity score where SLE uses one order higher than suggested by cross-validation. Standard errors (‘se’) are based on bootstrapping the effects 1999 times. P-values (‘pval’) of most estimators are computed according to the t-statistic, with the exception of ‘simulation’ and ‘simulation gam’, for which the p-values follow from simulating potential mediators and outcomes (1999 times) using the ‘mediation’ package for R of Tingley, Yamamoto, Hirose, Imai, and Keele (2014).<sup>7</sup>

None of the direct effects is statistically different from zero at conventional levels of significance, no matter which method is considered. All of the indirect effects are positive, amounting to an increase in (intended) circumcision of roughly one percentage point in most estimations. Even though the effects are small, many of them are statistically significant at the 10% level. This is the case for IPW using SLE with cross-validation or probit for propensity score estimation, g-computation, and multiply robust estimation. The indirect effects under simulation-based estimation and IPW with overfitted propensity scores are not significant at conventional levels, but the point estimates are similar to those of the other procedures. All in all, the results suggest that the information intervention has a very moderate positive effect on the outcome through a change in the assessment of HIV risk with and without circumcision. In contrast, the insignificant direct effects do not point to further important causal mechanisms through which information affects (intended) circumcision.

<sup>7</sup>Note that standard errors are not provided for the simulation-based estimators.

## 7 Conclusion

This paper proposed a fully nonparametric estimator of natural direct and indirect effects in the total population or specific subgroups based on inverse probability weighting and series logit estimation of the propensity scores when invoking a sequential conditional independence assumption. We established the conditions required for the root-n consistency and asymptotic normality of our estimator and investigated its finite sample performance in a simulation study. Finally, we applied our method to experimental data from Malawi to evaluate the direct effect of receiving information on circumcision on the inclination towards circumcision, as well as the indirect effect mediated by the risk assessment of HIV with and without circumcision. The results suggest a very moderate positive indirect effect while the direct effect is virtually zero and not statistically significant.

## References

- ALBERT, J. M. (2008): “Mediation analysis via potential outcomes models,” *Statistics in Medicine*, 27, 1282–1304.
- ALBERT, J. M., AND S. NELSON (2011): “Generalized causal mediation analysis,” *Biometrics*, 67, 1028–1038.
- AN, W., AND X. WANG (2016): “Instrumental Variable Estimation of Causal Effects through Local Average Response Functions,” *Journal of Statistical Software*, 71, 1–13.
- BARON, R. M., AND D. A. KENNY (1986): “The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations,” *Journal of Personality and Social Psychology*, 51, 1173–1182.
- BUSSO, M., J. DINARDO, AND J. MCCRARY (2014): “New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators,” *forthcoming in the Review of Economics and Statistics*.
- CHINKHUMBA, J., S. GODLONTON, AND R. THORNTON (2014): “The Demand for Medical Male Circumcision,” *American Economic Journal: Applied Economics*, 6, 152–177.
- DONALD, S. G., AND Y.-C. HSU (2014): “Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models,” *Journal of Econometrics*, 178, 383–397.
- DONALD, S. G., Y.-C. HSU, AND R. P. LIELI (2014): “Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT,” *Journal of Business & Economic Statistics*, 32(3), 395–415.
- FLORES, C. A., AND A. FLORES-LAGUNES (2009): “Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness,” *IZA Discussion Paper No. 4237*.



- GELMAN, A., AND G. IMBENS (2013): “Why ask Why? Forward Causal Inference and Reverse Causal Questions,” *NBER Working Paper No. 19614*.
- HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161–1189.
- HONG, G. (2010): “Ratio of mediator probability weighting for estimating natural direct and indirect effects,” in *JSM Proceedings, Biometrics Section*, pp. 2401–2415. American Statistical Association, Alexandria, VA.
- HSU, Y.-C. (2016): “Consistent Tests for Conditional Treatment Effects,” *working paper, Academia Sinica, Taipei*.
- HUBER, M. (2014): “Identifying causal mechanisms (primarily) based on inverse probability weighting,” *Journal of Applied Econometrics*, 29, 920–943.
- IMAI, K., L. KEELE, AND T. YAMAMOTO (2010): “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects,” *Statistical Science*, 25, 51–71.
- IMAI, K., AND T. YAMAMOTO (2013): “Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments,” *Political Analysis*, 21, 141–171.
- 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 J. M. WOOLDRIDGE (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.
- JUDD, C. M., AND D. A. KENNY (1981): “Process Analysis: Estimating Mediation in Treatment Evaluations,” *Evaluation Review*, 5, 602–619.
- PEARL, J. (2001): “Direct and indirect effects,” in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.
- PETERSEN, M. L., S. E. SINISI, AND M. J. VAN DER LAAN (2006): “Estimation of Direct Causal Effects,” *Epidemiology*, 17, 276–284.
- ROBINS, J. M. (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.
- (2003): “Semantics of causal DAG models and the identification of direct and indirect effects,” in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.
- ROBINS, J. M., AND S. GREENLAND (1992): “Identifiability and Exchangeability for Direct and Indirect Effects,” *Epidemiology*, 3, 143–155.
- RUBIN, D. B. (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.
- (2004): “Direct and Indirect Causal Effects via Potential Outcomes,” *Scandinavian Journal of Statistics*, 31, 161–170.
- TCHETGEN TCHETGEN, E. J. (2013): “Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis,” *Statistics in Medicine*, 32, 4567–4580.

- TCHETGEN TCHETGEN, E. J., AND I. SHPITSER (2012): “Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis,” *The Annals of Statistics*, 40, 1816–1845.
- TEN HAVE, T. R., M. M. JOFFE, K. G. LYNCH, G. K. BROWN, S. A. MAISTO, AND A. T. BECK (2007): “Causal mediation analyses with rank preserving models,” *Biometrics*, 63, 926–934.
- TINGLEY, D., T. YAMAMOTO, K. HIROSE, K. IMAI, AND L. KEELE (2014): “Mediation: R package for causal mediation analysis,” *Journal of Statistical Software*, 59, 1–38.
- VANDERWEELE, T. J. (2009): “Marginal Structural Models for the Estimation of Direct and Indirect Effects,” *Epidemiology*, 20, 18–26.
- VANSTEELANDT, S., M. BEKAERT, AND T. LANGE (2012): “Imputation Strategies for the Estimation of Natural Direct and Indirect Effects,” *Epidemiologic Methods*, 1, 129–158.
- ZHENG, W., AND M. J. VAN DER LAAN (2012): “Targeted Maximum Likelihood Estimation of Natural Direct Effects,” *The International Journal of Biostatistics*, 8, 1–40.

## A Appendix

### A.1 Proof of Theorem 1

Define  $Y_{dd'} = Y(d, M(d'))$  and  $\mu_{dd'} = E[Y_{dd'}]$  for  $d, d' \in \{0, 1\}$ . By Assumptions 1 and 2,

$$\begin{aligned}\mu_{11} &= E\left[\frac{DY}{p(X)}\right], & \mu_{00} &= E\left[\frac{(1-D)Y}{1-p(X)}\right], \\ \mu_{10} &= E\left[\frac{DY}{p(M, X)} \frac{1-p(M, X)}{1-p(X)}\right], & \mu_{01} &= E\left[\frac{(1-D)Y}{1-p(M, X)} \frac{p(M, X)}{p(X)}\right],\end{aligned}$$

see equations (4) and (5) of Huber (2014). This allows defining the direct and indirect effects of interest in terms of  $\mu_{dd'}$ , e.g.  $\theta(1) = \mu_{11} - \mu_{01}$ . We estimate  $\mu_{dd'}$  for  $d, d' \in \{0, 1\}$  by the normalized sample analogs

$$\begin{aligned}\hat{\mu}_{11} &= \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\hat{p}(X_i)}, & \hat{\mu}_{00} &= \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{10} &= \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n \frac{D_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{01} &= \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)}.\end{aligned}$$

To prove Theorem 1, it is thus sufficient to show that for  $d, d' \in \{0, 1\}$ ,

$$\sqrt{n}(\hat{\mu}_{dd'} - \mu_{dd'}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{dd'}(Y_i, M_i, D_i, X_i) + o_p(1).$$

As the results regarding  $\mu_{11}$  and  $\mu_{00}$  have been established by HIR, we subsequently focus on the proof for  $\mu_{10}$  and note that the derivations for  $\mu_{01}$  proceed in an analogous way. Let  $\tilde{\mu}_{10}$  be the numerator of  $\hat{\mu}_{10}$  and  $\tilde{\omega}_{10}$  be the denominator of  $\hat{\mu}_{10}$ . We first show that

$$\sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{10}(Y_i, M_i, D_i, X_i) + o_p(1),$$

using a similar approach as HIR. Note that

$$\begin{aligned}
\sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1 - \hat{p}(M_i, X_i)}{1 - \hat{p}(X_i)} - \mu_{10} \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{p(M_i, X_i)} \frac{1 - p(M_i, X_i)}{1 - p(X_i)} - \mu_{10} \right\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{p(M_i, X_i)} \frac{1 - p(M_i, X_i)}{(1 - p(X_i))^2} (\hat{p}(X_i) - p(X_i)) \right\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{-D_i Y_i}{p^2(M_i, X_i)} \frac{1}{1 - p(X_i)} (\hat{p}(M_i, X_i) - p(M_i, X_i)) \right\} + o_p(1)
\end{aligned}$$

where the second equality holds by a second order mean valued expansion around  $p(X_i)$  and  $p(M_i, X_i)$  and the fact that  $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$  and  $\sup_{m \in \mathcal{M}, x \in \mathcal{X}} |\hat{p}(m, x) - p(m, x)| = o_p(n^{-1/4})$ . The converge rate of  $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = O_p(K_x(\sqrt{K_x/n} + K^{-\bar{p}_x/2d_x}))$  is derived in Lemma A3 of Hsu (2016) and the conditions given in this paper are sufficient for  $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$ . By the same argument as in Theorem 1 of HIR, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{p(M_i, X_i)} \frac{1 - p(M_i, X_i)}{(1 - p(X_i))^2} (\hat{p}(X_i) - p(X_i)) \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ E \left[ \frac{D_i Y_i}{p(M_i, X_i)} \frac{1 - p(M_i, X_i)}{(1 - p(X_i))^2} \middle| X_i \right] (D_i - p(X_i)) \right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\rho_{10}(X_i)}{(1 - p(X_i))} (D_i - p(X_i)) \right\} + o_p(1).
\end{aligned}$$

Similarly, it holds that

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{-D_i Y_i}{p^2(M_i, X_i)} \frac{1}{1 - p(X_i)} (\hat{p}(M_i, X_i) - p(M_i, X_i)) \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ E \left[ \frac{-D_i Y_i}{p^2(M_i, X_i)} \frac{1}{1 - p(X_i)} \middle| M_i, X_i \right] (D_i - p(M_i, X_i)) \right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{-\zeta_1(M_i, X_i)}{p(M_i, X_i)} \frac{1}{1 - p(X_i)} (D_i - p(M_i, X_i)) \right\} + o_p(1).
\end{aligned}$$

Next, one can easily show that

$$\sqrt{n}(\tilde{w}_{10} - 1) = o_p(1),$$

by replacing  $Y_i$ 's with 1's in the proof for  $\tilde{\mu}_{10}$  and acknowledging that the influence functions for

$\tilde{w}_{10}$  are all zero. Finally, it follows that

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{10} - \mu_{10}) &= \sqrt{n}\left(\frac{\tilde{\mu}_{10}}{\tilde{w}_{10}} - \mu_{10}\right) \\ &= \sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) - \mu_{10}\sqrt{n}(\tilde{w}_{10} - 1) + o_p(1) \\ &= \sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) + o_p(1),\end{aligned}$$

where the second equality follows by a mean-value expansion and the last equality holds by the fact that  $\mu_{10}$  is bounded and that  $\sqrt{n}(\tilde{w}_{10} - 1) = o_p(1)$ . This completes the proof.  $\square$

## A.2 Proof of Theorem 2

Define  $\mu_{dd',g} = E[g(X)Y_{dd'}]/E[g(X)]$  for  $d, d' \in \{0, 1\}$ . Similarly,

$$\begin{aligned}\mu_{11,g} &= \frac{E\left[g(X)\frac{DY}{p(X)}\right]}{E[g(X)]}, \\ \mu_{00,g} &= \frac{E\left[g(X)\frac{(1-D)Y}{1-p(X)}\right]}{E[g(X)]}, \\ \mu_{10,g} &= \frac{E\left[g(X)\frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)}\right]}{E[g(X)]}, \\ \mu_{01,g} &= \frac{E\left[g(X)\frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)}\right]}{E[g(X)]}.\end{aligned}$$

We estimate  $\mu_{dd',g}$  for  $d, d' \in \{0, 1\}$  by the normalized sample analogs

$$\begin{aligned}\hat{\mu}_{11,g} &= \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i Y_i}{\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i}{\hat{p}(X_i)}, \\ \hat{\mu}_{00,g} &= \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{(1-D_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{10,g} &= \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{01,g} &= \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{(1-D_i)Y_i}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \bigg/ \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{(1-D_i)}{1-\hat{p}(M_i, X_i)} \frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)}.\end{aligned}$$

Define  $P_g = E[g(X)]$ . We would like to show that  $\sqrt{n}(\hat{\mu}_{11,g} - \mu_{11,g}) = n^{-1/2} \sum_{i=1}^n \psi_{\mu_{11,g}}(Y_i, M_i, D_i, X_i) + o_p(1)$ . First, we can be demonstrated that

$$\begin{aligned}
& \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i Y_i}{\hat{p}(X_i)} - \mu_{11,g} \cdot P_g \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{g(X_i) D_i Y_i}{p(X_i)} - \frac{g(X_i) \rho_{11}(X_i)}{p(X_i)} (D_i - p(X_i)) - \mu_{11,g} \cdot P_g \right) + o_p(1), \text{ and} \\
& \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n g(X_i) \frac{D_i}{\hat{p}(X_i)} - \mu_{11,g} \cdot P_g \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{g(X_i) D_i}{p(X_i)} - \frac{g(X_i)}{p(X_i)} (D_i - p(X_i)) - P_g \right) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(X_i) - P_g) + o_p(1).
\end{aligned}$$

Then by delta method, we have

$$\begin{aligned}
& \sqrt{n}(\hat{\mu}_{11,g} - \mu_{11,g}) \\
&= \frac{1}{P_g} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{g(X_i) D_i Y_i}{p(X_i)} - \frac{g(X_i) \mu_1(X_i)}{p(X_i)} (D_i - p(X_i)) - \mu_{11,g} \cdot P_g \right) - \frac{\mu_{11,g}}{P_g} (g(X_i) - P_g) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{g(X_i)}{E[g(X)]} \left( \frac{D_i Y_i}{p(X_i)} - \frac{\rho_{11}(X_i)}{p(X_i)} (D_i - p(X_i)) - \mu_{11,g} \right) + o_p(1).
\end{aligned}$$

This gives the result for the  $\mu_{11,g}$  case. Using the same arguments, we can derive the results for  $\mu_{00,g}$ ,  $\mu_{01,g}$  and  $\mu_{10,g}$ , too, which is sufficient to prove Theorem 2.  $\square$

### A.3 Proof of Theorem 3

Define  $\mu_{dd',t} = E[p(X)Y_{dd'}]/E[p(X)]$  for  $d, d' \in \{0, 1\}$ . Similarly,

$$\begin{aligned}
\mu_{11,t} &= \frac{E \left[ p(X) \frac{DY}{p(X)} \right]}{E[p(X)]}, \\
\mu_{00,t} &= \frac{E \left[ p(X) \frac{(1-D)Y}{1-p(X)} \right]}{E[p(X)]}, \\
\mu_{10,t} &= \frac{E \left[ p(X) \frac{DY}{p(M,X)} \frac{1-p(M,X)}{1-p(X)} \right]}{E[p(X)]}, \\
\mu_{01,t} &= \frac{E \left[ p(X) \frac{(1-D)Y}{1-p(M,X)} \frac{p(M,X)}{p(X)} \right]}{E[p(X)]}.
\end{aligned}$$

We estimate  $\mu_{dd',t}$  for  $d, d' \in \{0, 1\}$  by the normalized sample analogs

$$\begin{aligned}\hat{\mu}_{11,t} &= \frac{1}{n} \sum_{i=1}^n D_i Y_i \Big/ \frac{1}{n} \sum_{i=1}^n D_i, \\ \hat{\mu}_{00,t} &= \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{(1-D_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{10,t} &= \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{D_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{01,t} &= \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)Y_i}{1-\hat{p}(M_i, X_i)} \hat{p}(M_i, X_i) \Big/ \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(M_i, X_i)} \hat{p}(M_i, X_i).\end{aligned}$$

Define  $P_t = E[p(X)]$ . Note that we have

$$\begin{aligned}\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n D_i Y_i - \mu_{11,t} P_t \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i Y_i - \mu_{11,t} P_t), \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - \mu_{00,t} P_t \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{p(X_i)(1-D_i)Y_i}{1-p(X_i)} + \frac{\mu_0(X_i)}{1-p(X_i)} (D_i - p(X_i)) - \mu_{00,t} P_t \right) + o_p(1), \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{D_i Y_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} - \mu_{10,t} P_t \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{p(X_i) D_i Y_i}{p(M_i, X_i)} \frac{1-p(M_i, X_i)}{1-p(X_i)} + \frac{\rho_{10}(X_i)}{1-p(X_i)} (D_i - p(X_i)) \right. \\ &\quad \left. - \frac{p(X_i) \zeta_1(M_i, X_i)}{p(M_i, X_i)(1-p(X_i))} (D_i - p(M_i, X_i)) - \mu_{10,t} P_t \right) + o_p(1), \text{ and} \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)Y_i}{1-\hat{p}(M_i, X_i)} \hat{p}(M_i, X_i) - \mu_{01,t} P_t \right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{(1-D_i)Y_i}{1-p(M_i, X_i)} p(M_i, X_i) + \frac{\zeta_0(M_i, X_i)}{(1-p(M_i, X_i))} (D_i - p(M_i, X_i)) - \mu_{01,t} P_t \right) + o_p(1).\end{aligned}$$

Furthermore,

$$\begin{aligned}\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n D_i - P_t \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - P_t) + o_p(1), \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - P_t \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - P_t) + o_p(1) \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \hat{p}(X_i) \frac{D_i}{\hat{p}(M_i, X_i)} \frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} - P_t \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - P_t) + o_p(1), \text{ and} \\ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i)}{1-\hat{p}(M_i, X_i)} \hat{p}(M_i, X_i) - P_t \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - P_t) + o_p(1).\end{aligned}$$

By delta method, we have

$$\begin{aligned}
& \sqrt{n}(\hat{\mu}_{11,t} - \mu_{11,t}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( D_i Y_i - \mu_{11,t} P_t - \mu_{11,t} (D_i - P_t) \right) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( D_i (Y_i - \mu_{11,t}) \right) + o_p(1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \sqrt{n}(\hat{\mu}_{00,t} - \mu_{00,t}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( \frac{p(X_i)(1-D_i)Y_i}{1-p(X_i)} + \frac{\rho_{00}(X_i)}{1-p(X_i)} (D_i - p(X_i)) - \mu_{00,t} P_t - \mu_{00,t} (D_i - P_t) \right) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( \frac{p(X_i)(1-D_i)(Y_i - \rho_{00}(X_i))}{1-p(X_i)} + (\rho_{00}(X_i) - \mu_{00,t}) D_i \right) + o_p(1).
\end{aligned}$$

Next,

$$\begin{aligned}
& \sqrt{n}(\hat{\mu}_{10,t} - \mu_{10,t}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( \frac{p(X_i) D_i Y_i}{p(M_i, X_i)} \frac{1-p(M_i, X_i)}{1-p(X_i)} + \frac{\rho_{10}(X_i)}{(1-p(X_i))} (D_i - p(X_i)) \right. \\
&\quad \left. - \frac{p(X_i) \zeta_1(M_i, X_i)}{p(M_i, X_i)(1-p(X_i))} (D_i - p(M_i, X_i)) - \mu_{10,t} D_i \right) + o_p(1), \\
& \sqrt{n}(\hat{\mu}_{01,t} - \mu_{01,t}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{P_t} \left( \frac{(1-D_i)Y_i}{1-p(M_i, X_i)} p(M_i, X_i) + \frac{\zeta_0(M_i, X_i)}{(1-p(M_i, X_i))} (D_i - p(M_i, X_i)) - \mu_{01,t} D_i \right) + o_p(1).
\end{aligned}$$

This completes the proof.  $\square$