

# Identification and Inference for Welfare Gains without Unconfoundedness \*

Job Market Paper

Undral Byambadalai †

January 26, 2021

[please click here for the latest version](#)

## Abstract

This paper studies identification and inference of the welfare gain that results from switching from one policy (such as the status quo policy) to another policy. The welfare gain is not point identified in general when data are obtained from an observational study or a randomized experiment with imperfect compliance. I characterize the sharp identified region of the welfare gain and obtain bounds under various assumptions on the unobservables with and without instrumental variables. Estimation and inference of the lower and upper bounds are conducted using orthogonalized moment conditions to deal with the presence of infinite-dimensional nuisance parameters. I illustrate the analysis by considering hypothetical policies of assigning individuals to job training programs using experimental data from the National Job Training Partnership Act Study. Monte Carlo simulations are conducted to assess the finite sample performance of the estimators.

**Keywords:** treatment assignment, observational data, partial identification, semiparametric inference

**JEL Classifications:** C01, C13, C14

---

\*I am deeply indebted to my main adviser Hiroaki Kaido for his unparalleled guidance and constant encouragement throughout this project. I am grateful to Iván Fernández-Val and Jean-Jacques Forneron for their invaluable suggestions and continuous support. For their helpful comments and discussions, I thank Karun Adusumilli, Fatima Aqeel, Susan Athey, Jesse Bruhn, Alessandro Casini, Mingli Chen, Shuowen Chen, David Childers, Taosong Deng, Wayne Gao, Thea How Choon, David Kaplan, Shakeeb Khan, Louise Laage, Michael Leung, Arthur Lewbel, Jessie Li, Jia Li, Siyi Luo, Ye Luo, Ching-to Albert Ma, Francesca Molinari, Guillaume Pouliot, Patrick Power, Anlong Qin, Zhongjun Qu, Enrique Sentana, Vasilis Syrgkanis, Purevdorj Tuvaandorj, Guang Zhang, and seminar participants at Boston University and participants of BU-BC Joint Workshop in Econometrics 2019, the Econometric Society/Bocconi University World Congress 2020, and the European Economic Association Annual Congress 2020. All errors are my own and comments are welcome.

†Department of Economics, Boston University, 270 Bay State Rd., Boston, MA 02215 Email: [undral@bu.edu](mailto:undral@bu.edu)  
Website: [undralbyambadalai.com](http://undralbyambadalai.com)

# 1 Introduction

The problem of choosing among alternative treatment assignment rules based on data is pervasive in economics and many other fields, including marketing and medicine. A treatment assignment rule is a mapping from individual characteristics to a treatment assignment. For instance, it can be a job training program eligibility criterion based on the applicants' years of education and annual earnings. Throughout the paper, I call the treatment assignment rule a policy, and the subject who decides the treatment assignment rule a policymaker. The policymaker can be an algorithm assigning targeted ads, a doctor deciding medical treatment, or a school principal deciding which students take classes in person during a pandemic. As individuals with different characteristics might respond differently to a given policy, policymakers aim to choose a policy that generates the highest overall outcome or welfare.

Most previous work on treatment assignment in econometrics focused on estimating the optimal policy using data from a randomized experiment. I contribute to this literature by focusing on the identification and inference of the *welfare gain* using data from an observational study or a randomized experiment with imperfect compliance. The assumption called *unconfoundedness* might fail to hold for such datasets.<sup>1</sup> By relaxing the unconfoundedness assumption, my framework accommodates many interesting and empirically relevant cases, including the use of instrumental variables to identify the effect of a treatment. The advantage of focusing on welfare gain is to provide policymakers with the ability to be more transparent when choosing among alternative policies. Policymakers may want to know how much the welfare gain or loss is in addition to the welfare ranking of competing policies when they make their decisions. They might also need to report the welfare gain.

When the unconfoundedness assumption does not hold, identification of the conditional average treatment effect (CATE) and hence identification of the welfare gain becomes a delicate matter. Without further assumptions on selection, one cannot uniquely identify the welfare gain. I take a partial identification approach whereby one obtains bounds on the parameter of interest with a minimal amount of assumptions on the unobservables and, later on, tighten these bounds by imposing additional assumptions with and without instrumental variables. The bounds, or sharp identified region, of the welfare gain can be characterized using tools from random set theory.<sup>2</sup> The framework I use allows me to consider various assumptions that involve instrumental variables and shape restrictions on the unobservables.

---

<sup>1</sup> The assumption of unconfoundedness is also known as *selection on observables* and assumes that treatment is independent of potential outcomes conditional on observable characteristics.

<sup>2</sup> The terms *identified region*, *identified set*, and *bounds* are used interchangeably throughout the paper. Often the word *sharp* is omitted, and unless explicitly described as non-sharp, identified region/identified set/bounds refer to sharp identified region/sharp identified set/sharp bounds.

I show that the lower and upper bounds of the welfare gain can, in general, be written as functions of the conditional mean treatment responses and a propensity score. Hence, estimation and inference of these bounds can be thought of as a semiparametric estimation problem in which the conditional mean treatment responses and the propensity score are infinite-dimensional nuisance parameters. Bounds that do not rely on instruments admit regular and asymptotically normal estimators. I construct orthogonalized, or locally robust, moment condition by adding an adjustment term that accounts for the first-step estimation to the original moment condition, following [Chernozhukov, Escanciano, Ichimura, Newey, and Robins \(2020\)](#) (CEINR, henceforth). This method leads to estimators that are first-order insensitive to estimation errors of the nuisance parameters. I calculate the adjustment term using an approach proposed by [Ichimura and Newey \(2017\)](#). The locally robust estimation is possible even with instrumental variables under an additional monotonicity assumption of instruments. The estimation strategy has at least two advantages. First, it allows for flexible estimation of nuisance parameters, including the possibility of using high-dimensional machine learning methods. Second, the calculation of confidence intervals for the bounds is straightforward because the asymptotic variance doesn't rely on the estimation of nuisance parameters.

I illustrate the analysis using experimental data from the National Job Training Partnership Act (JTPA) Study. This dataset has been analyzed extensively in economics to understand the effect of subsidized training on outcomes such as earnings. I consider two hypothetical examples. First, I compare two different treatment assignment policies that are functions of individuals' years of education. Second, I compare [Kitagawa and Tetenov \(2018\)](#)'s estimated optimal policy with an alternative policy when the conditioning variables are individuals' years of education and pre-program annual earnings. The results from a Monte Carlo simulation suggest that the method works well in a finite sample.

## 1.1 Related Literature

This paper is related to the literature on treatment assignment, sometimes also referred to as treatment choice, which has been growing in econometrics since the seminal work by [Manski \(2004\)](#). Earlier work in this literature include [Dehejia \(2005\)](#), [Hirano and Porter \(2009\)](#), [Stoye \(2009a, 2012\)](#), [Chamberlain \(2011\)](#), [Bhattacharya and Dupas \(2012\)](#), [Tetenov \(2012\)](#), [Kasy \(2014\)](#), and [Armstrong and Shen \(2015\)](#).

In a recent work, [Kitagawa and Tetenov \(2018\)](#) propose what they call an empirical welfare maximization method. This method selects a treatment rule that maximizes the sample analog of the average social welfare over a class of candidate treatment rules. Their method has been further studied and extended in different directions. [Kitagawa and Tetenov \(2019\)](#) study

an alternative welfare criterion that concerns equality. [Mbakop and Tabord-Meehan \(2016\)](#) propose what they call a penalized welfare maximization, an alternative method to estimate optimal treatment rules. While [Andrews, Kitagawa, and McCloskey \(2019\)](#) consider inference for the estimated optimal rule, [Rai \(2018\)](#) considers inference for the optimal rule itself. These papers and most of the earlier papers only apply to a setting in which the assumption of unconfoundedness holds.

In a dynamic setting, treatment assignment is studied by [Kock and Thyrgaard \(2017\)](#), [Kock, Preinerstorfer, and Veliyev \(2018\)](#), [Adusumilli, Geiecke, and Schilter \(2019\)](#), [Sakaguchi \(2019\)](#), and [Han \(2019\)](#), among others.

This paper contributes to the less explored case of using observational data to infer policy choice where the unconfoundedness assumption does not hold. Earlier work in the treatment choice literature with partial identification include [Stoye \(2007\)](#) and [Stoye \(2009b\)](#). This paper is closely related to [Kasy \(2016\)](#), but their main object of interest is the welfare ranking of policies rather than the magnitude of welfare gain that results from switching from one policy to another policy. It is also closely related to [Athey and Wager \(2020\)](#) as they are concerned with choosing treatment assignment policies using observational data. However, their approach is about estimating the optimal treatment rule by point identifying the causal effect using various assumptions. In a related work in statistics, [Cui and Tchetgen Tchetgen \(2020\)](#) propose a method to estimate optimal treatment rules using instrumental variables. More recently, [Assunção, McMillan, Murphy, and Souza-Rodrigues \(2019\)](#) work with a partially identified welfare criterion that also takes spillover effects into account to analyze deforestation regulations in Brazil.

The rest of the paper is structured as follows. In Section 2, I set up the problem. Section 3 presents the identification results of the welfare gain. Section 4 discusses the estimation and inference of the bounds. In Section 5, I illustrate the analysis using experimental data from the National JTPA study. Section 6 summarizes the results from a Monte Carlo simulation. Finally, Section 7 concludes. All proofs, some useful definitions and theorems from random set theory, additional tables and figures from the empirical application, and more details on the simulation study are collected in the Appendix.

**Notation.** Throughout the paper, for  $d \in \mathbb{N}$ , let  $\mathbb{R}^d$  denote the Euclidean space and  $\|\cdot\|$  denote the Euclidean norm. Let  $\langle \cdot, \cdot \rangle$  denote the inner product in  $\mathbb{R}^d$  and  $E[\cdot]$  denote the expectation operator. The notation  $\xrightarrow{p}$  and  $\xrightarrow{d}$  denote convergence in probability and convergence in distribution, respectively. For a sequence of numbers  $x_n$  and  $y_n$ ,  $x_n = o(y_n)$  and  $x_n = O(y_n)$  mean, respectively, that  $x_n/y_n \rightarrow 0$  and  $x_n \leq Cy_n$  for some constant  $C$  as  $n \rightarrow \infty$ . For a sequence of random variables  $X_n$  and  $Y_n$ , the notation  $X_n = o_p(Y_n)$  and  $X_n = O_p(Y_n)$  mean,

respectively, that  $X_n/Y_n \xrightarrow{p} 0$  and  $X_n/Y_n$  is bounded in probability.  $\mathcal{N}(\mu, \Omega)$  denotes a normal distribution with mean  $\mu$  and variance  $\Omega$ .  $\Phi(\cdot)$  denotes the cumulative distribution function of the standard normal distribution.

## 2 Setup

Let  $(\Omega, \mathfrak{A})$  be a measurable space. Let  $Y : \Omega \rightarrow \mathbb{R}$  denote an outcome variable,  $D : \Omega \rightarrow \{0, 1\}$  denote a binary treatment, and  $X : \Omega \rightarrow \mathcal{X} \subset \mathbb{R}^{d_x}$  denote pretreatment covariates. For  $d \in \{0, 1\}$ , let  $Y_d : \Omega \rightarrow \mathbb{R}$  denote a potential outcome that would have been observed if the treatment status were  $D = d$ . For each individual, the researcher only observes either  $Y_1$  or  $Y_0$  depending on what treatment the individual received. Hence, the relationship between observed and potential outcomes is given by

$$Y = Y_1 \cdot D + Y_0 \cdot (1 - D). \quad (1)$$

Policy I consider is a treatment assignment rule based on observed characteristics of individuals. In other words, the policymaker assigns an individual with covariate  $X$  to a binary treatment according to a treatment rule  $\delta : \mathcal{X} \rightarrow \{0, 1\}$ .<sup>3</sup> The welfare criterion considered is population mean welfare. If the policymaker chooses policy  $\delta$ , the welfare is given by

$$\begin{aligned} u(\delta) &\equiv E[Y_1 \cdot \delta(X) + Y_0 \cdot (1 - \delta(X))] \\ &= E[E[Y_1|X] \cdot \delta(X) + E[Y_0|X] \cdot (1 - \delta(X))]. \end{aligned} \quad (2)$$

The object of my interest is *welfare gain* that results from switching from policy  $\delta^*$  to another policy  $\delta$  which is

$$u(\delta) - u(\delta^*) = E[\Delta(X) \cdot (\delta(X) - \delta^*(X))], \quad \Delta(X) \equiv E[Y_1 - Y_0|X]. \quad (3)$$

**Remark 1.** I assume that individuals comply with the assignment. This can serve as a natural baseline for choosing between policies.

The observable variables in my model are  $(Y, D, X)$  and I assume that the researcher knows the joint distribution of  $(Y, D, X)$  when I study identification. Later, in Section 4, I assume availability of data – size  $n$  random sample from  $(Y, D, X)$  – to conduct inference on objects that depend on this joint distribution. The unobservables in my model are potential outcomes

---

<sup>3</sup>I consider deterministic treatment rules in my framework. See Appendix C for discussions on randomized treatment rules.

$(Y_1, Y_0)$ . The conditional average treatment effect  $\Delta(X) = E[Y_1 - Y_0|X]$  and hence my object of interest welfare gain cannot be point identified in the absence of strong assumptions. One instance in which it can be point identified is when potential outcomes  $(Y_1, Y_0)$  are independent of treatment  $D$  conditional on  $X$ , i.e.,

$$(Y_1, Y_0) \perp D|X. \quad (4)$$

This assumption is called *unconfoundedness* and is a widely-used identifying assumption in causal inference. See [Imbens and Rubin \(2015\)](#) Chapter 12 and 21 for more discussions on this assumption. Under unconfoundedness, the conditional average treatment effect can be identified as

$$E[Y_1 - Y_0|X] = E[Y|D = 1, X] - E[Y|D = 0, X]. \quad (5)$$

Note that the right-hand side of (5) is identified since the researcher knows the joint distribution of  $(Y, D, X)$ . If data are obtained from a randomized experiment, the assumption holds since the treatment is randomly assigned. However, if data are obtained from an observational study, the assumption is not testable and often controversial. In the next section, I relax the assumption of unconfoundedness and explore what can be learned about my parameter of interest when different assumptions are imposed on the unobservables and when there are additional instrumental variables  $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$  to help identify the conditional average treatment effect.

The welfare gain is related to [Manski \(2004\)](#)'s *regret* which has been used by [Kitagawa and Tetenov \(2018\)](#), [Athey and Wager \(2020\)](#), and many others in the literature to evaluate the performance of the estimated treatment rules. When  $\mathcal{D}$  is the class of treatment rules to be considered, the regret from choosing treatment rule  $\delta$  is  $u(\delta^*) - u(\delta)$  where

$$\delta^* = \arg \max_{d \in \mathcal{D}} E[E[Y_1|X] \cdot d + E[Y_0|X] \cdot (1 - d)]. \quad (6)$$

It is an expected loss in welfare that results from not reaching the maximum feasible welfare as  $\delta^*$  is the policy that maximizes population welfare. In [Kitagawa and Tetenov \(2018\)](#) and others, under the assumption of unconfoundedness, the welfare criterion  $u(\delta)$  in (2) is point-identified. Therefore, the optimal "oracle" treatment rule in (6) is well defined when the researcher knows the joint distribution of  $(Y, D, X)$ . However, when the welfare criterion in (2) is set-identified, one needs to specify their notion of optimality. For instance, the optimal rule could be a rule that maximizes the guaranteed or minimum welfare.

## 3 Identification

### 3.1 Sharp identified region

Partial identification approach has been proven to be a useful alternative or complement to point identification analysis with strong assumptions. See [Manski \(2003\)](#), [Tamer \(2010\)](#), and [Molinari \(2019\)](#) for an overview. The theory of random sets, which I use to conduct my identification analysis, is one of the tools that have been used fruitfully to address identification and inference in partially identified models. Examples include [Beresteanu and Molinari \(2008\)](#), [Beresteanu, Molchanov, and Molinari \(2011, 2012\)](#), [Galichon and Henry \(2011\)](#), [Epstein, Kaido, and Seo \(2016\)](#), [Chesher and Rosen \(2017\)](#), and [Kaido and Zhang \(2019\)](#). See [Molchanov and Molinari \(2018\)](#) for a textbook treatment of its use in econometrics.

My goal in this section is to characterize the *sharp identified region* of the welfare gain when different assumptions are imposed on the unobservables. The sharp identified region of the welfare gain is the tightest possible set that collects the values of welfare gain that results from all possible  $(Y_1, Y_0)$  that are consistent with the maintained assumptions. Toward this end, I define a *random set* and its *selections* whose formal definitions can be found in [Appendix A](#). The random set is useful for incorporating weak assumptions in a unified framework rather than deriving bounds on a case-by-case basis. Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  be a random set where  $\mathcal{F}$  is the family of closed subsets of  $\mathbb{R}^2$ . Assumptions on potential outcomes can be imposed through this random set. Then, the collection of all random vectors  $(Y_1, Y_0)$  that are consistent with those assumptions equals the family of all selections of  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  denoted by  $\mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ . Specific examples of a random set with more discussions on selections, namely, in the context of worst-case bounds of [Manski \(1990\)](#) and monotone treatment response analysis of [Manski \(1997\)](#), are given in [Section 3.3](#). Using the random set notations I just introduced, the sharp identified region of the welfare gain is given by

$$B_I(\delta, \delta^*) \equiv \{\beta \in \mathbb{R} : \beta = E[E[Y_1 - Y_0|X] \cdot (\delta(X) - \delta^*(X))], (Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0)\}. \quad (7)$$

### 3.2 Lower and upper bound

One way to achieve characterization of the sharp identified region is through a *selection expectation* and its *support function*. Their definitions can be found in [Appendix A](#). Let the support function of a convex set  $K \subset \mathbb{R}^d$  be denoted by

$$s(v, K) = \sup_{x \in K} \langle v, x \rangle, \quad v \in \mathbb{R}^d. \quad (8)$$



The support function appears in Beresteanu and Molinari (2008), Beresteanu, Molchanov, and Molinari (2011), Bontemps, Magnac, and Maurin (2012), Kaido and Santos (2014), Kaido (2016), and Kaido (2017), among others.

I first state a lemma that will be useful to prove my main result. It shows how expectation of a functional of potential outcomes can be bounded from below and above by expected support function of the random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$ . The proof of the following lemma and all other proofs in this paper are collected in the Appendix.

**Lemma 1.** *Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  be an integrable random set that is almost surely convex and let  $(Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ . For any  $v \in \mathbb{R}^2$ , we have*

$$-E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] \leq v'E[(Y_1, Y_0)'|X] \leq E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] \quad a.s. \quad (9)$$

I introduce a notation that appears in the following theorem and throughout the paper. Let  $\theta_{10}(X) \equiv \mathbb{1}\{\delta(X) = 1, \delta^*(X) = 0\}$  be an indicator function for the sub population that are newly treated under the new policy. Similary, let  $\theta_{01}(X) \equiv \mathbb{1}\{\delta(X) = 0, \delta^*(X) = 1\}$  be an indicator function for the sub population that are no longer being treated because of the new policy.

**Theorem 1** (General case). *Suppose  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  is an integrable random set that is almost surely convex. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Also, let  $v^* = (1, -1)'$ . Then,  $B_I(\delta, \delta^*)$  in (7) is an interval  $[\beta_l, \beta_u]$  where*

$$\beta_l = E[\underline{\Delta}(X) \cdot \theta_{10}(X) - \bar{\Delta}(X) \cdot \theta_{01}(X)], \quad (10)$$

and

$$\beta_u = E[\bar{\Delta}(X) \cdot \theta_{10}(X) - \underline{\Delta}(X) \cdot \theta_{01}(X)], \quad (11)$$

where  $\underline{\Delta}(X) \equiv -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$  and  $\bar{\Delta}(X) \equiv E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$ .

The lower (upper) bound on the welfare gain is achieved when the newly treated people are the ones who benefit the least (most) from the treatment and the people who are no longer being treated are the ones who benefit the most (least) from the treatment. Therefore, the lower and upper bounds of the welfare gain involve both  $\underline{\Delta}(X) = -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$  and  $\bar{\Delta}(X) = E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$ , expected support functions of the random set at directions  $-v^* = (-1, 1)'$  and  $v^* = (1, -1)'$ . Oftentimes, these can be estimated by its sample analog estimators. I give closed form expressions of the expected support functions in Section 3.3 and 3.4 – they depend on objects such as  $E[Y|D = 1, X = x]$ ,  $E[Y|D = 0, X = x]$ , and



$P(D = 1|X = x)$ . To ease notation, let  $\eta(d, x) \equiv E[Y|D = d, X = x]$  for  $d \in \{0, 1\}$  be the conditional mean treatment responses and  $p(x) \equiv P(D = 1|X = x)$  be the propensity score.

While I characterize the identified region of the welfare gain directly given assumptions on the selections  $(Y_1, Y_0)$ , [Kasy \(2016\)](#)'s analysis is based on the identified set for CATE and their main results apply to any approach that leads to partial identification of treatment effects. The characterization I give above is related to their characterization when no restrictions across covariate values are imposed on treatment effects (e.g., no restrictions such as  $\Delta(x)$  is monotone in  $x$ ) and  $\underline{\Delta}(x)$  and  $\bar{\Delta}(x)$  are respectively lower and upper bound on the CATE  $\Delta(x)$ . As examples of such bounds, [Kasy \(2016\)](#) considers bounds that arise under instrument exogeneity as in [Manski \(2003\)](#) and under marginal stationarity of unobserved heterogeneity in panel data models as in [Chernozhukov, Fernández-Val, Hahn, and Newey \(2013\)](#). I consider bounds when there are instrumental variables that satisfy mean independence or mean monotonicity conditions as in [Manski \(2003\)](#) in Section 3.4.

In the following subsection, Section 3.3, I illustrate the form of the random set and show how Theorem 1 can be used to derive closed form bounds under different sets of assumptions.

### 3.3 Identification without Instruments

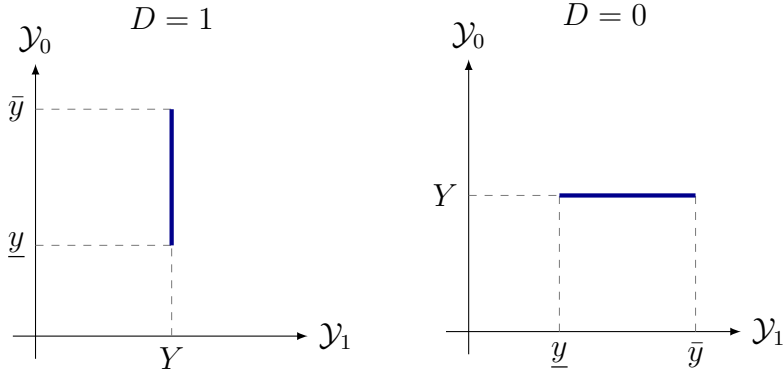


Figure 1: Random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  under worst-case

[Manski \(1990\)](#) derived worst-case bounds on  $Y_1$  and  $Y_0$  when the outcome variable is bounded, i.e.,  $Y \in [\underline{y}, \bar{y}] \subset \mathbb{R}$  where  $-\infty < \underline{y} \leq \bar{y} < \infty$ . It is called worst-case bounds because no additional assumptions are imposed on their distributions. Then, as shown in Figure 1, the

random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  is such that

$$\mathcal{Y}_1 \times \mathcal{Y}_0 = \begin{cases} \{Y\} \times [\underline{y}, \bar{y}] & \text{if } D = 1, \\ [\underline{y}, \bar{y}] \times \{Y\} & \text{if } D = 0. \end{cases} \quad (12)$$

The random set in (12) switches its value between two sets depending on the value of  $D$ . If  $D = 1$ ,  $\mathcal{Y}_1$  is given by a singleton  $\{Y\}$  whereas  $\mathcal{Y}_0$  is given by the entire support  $[\underline{y}, \bar{y}]$ . Similarly, if  $D = 0$ ,  $\mathcal{Y}_0$  is given by a singleton  $\{Y\}$  whereas  $\mathcal{Y}_1$  is given by the entire support  $[\underline{y}, \bar{y}]$ . I plot  $\mathcal{Y}_d$  and its selection  $Y_d$  for  $d \in \{0, 1\}$  as a function of  $\omega \in \Omega$  in Figure 2. If  $D = d$ , the random set  $\mathcal{Y}_d$  is a singleton  $\{Y\}$  and the family of selections consists of single random variable  $\{Y\}$  as well. On the other hand, if  $D = 1 - d$ , the random set  $\mathcal{Y}_d$  is an interval  $[\underline{y}, \bar{y}]$  and the family of all selections consists of all  $\mathfrak{A}$ -measurable random variables that has support on  $[\underline{y}, \bar{y}]$ . Note that each selection  $(Y_1, Y_0)$  of  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  can be represented in the following way. Take

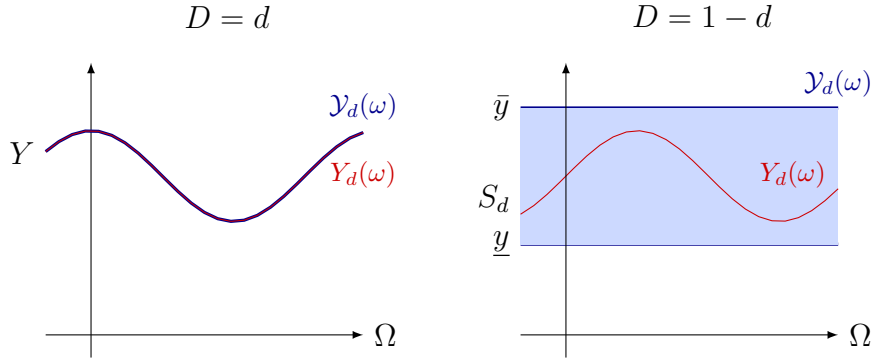


Figure 2: Random set  $\mathcal{Y}_d$  and its selection  $Y_d$  for  $d \in \{0, 1\}$  as a function of  $\omega \in \Omega$  under worst-case

random variables  $S_1 : \Omega \rightarrow \mathbb{R}$  and  $S_0 : \Omega \rightarrow \mathbb{R}$  whose distributions conditional on  $Y$  and  $D$  are not specified and can be any probability distributions on  $[\underline{y}, \bar{y}]$ . Then  $(Y_1, Y_0)$  that satisfies the following is a selection of  $\mathcal{Y}_1 \times \mathcal{Y}_0$ :

$$\begin{aligned} Y_1 &= Y \cdot D + S_1 \cdot (1 - D), \\ Y_0 &= Y \cdot (1 - D) + S_0 \cdot D. \end{aligned} \quad (13)$$

This representation makes it even clearer how I am not imposing any structure on the counterfactuals that I do not observe.  $S_1$  and  $S_0$  correspond to the *selection mechanisms* that appear in Ponomareva and Tamer (2011) and Tamer (2010).

Now, for the random set in (12), I can calculate its expected support function at directions

$v^* = (1, -1)$  and  $-v^* = (-1, 1)$  to obtain the bounds of the welfare gain in closed form. As shown in Figure 3, the support function of random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  in (12) at direction  $v^* = (1, -1)$  is the (signed) distance (rescaled by the norm of  $v^*$ ) between the origin and the hyperplane tangent to the random set in direction  $v^* = (1, -1)$ . Then, the bounds are given in the following Corollary to Theorem 1.

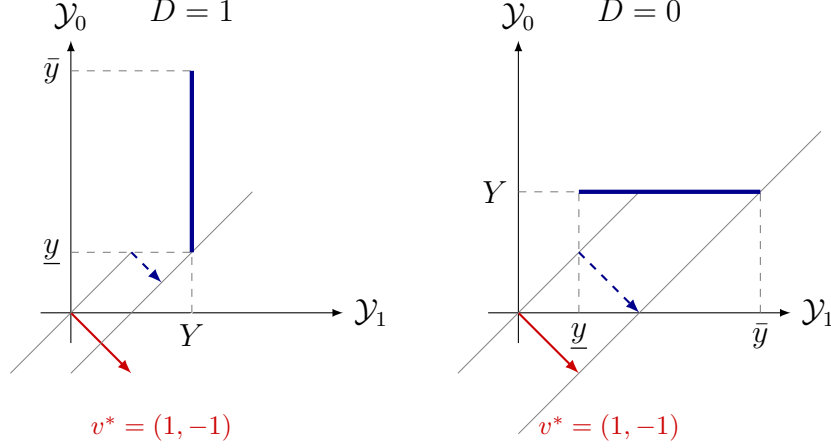


Figure 3: Support function of  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  at direction  $v^* = (1, -1)$  under worst-case

**Corollary 1** (Worst-case). *Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  be a random set in (12). Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Then,  $B_I(\delta, \delta^*)$  in (7) is an interval  $[\beta_l, \beta_u]$  where*

$$\begin{aligned} \beta_l = E[ & ((\eta(1, X) - \bar{y}) \cdot p(X) + (\underline{y} - \eta(0, X)) \cdot (1 - p(X))) \cdot \theta_{10}(X) \\ & - ((\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X))) \cdot \theta_{01}(X)], \end{aligned} \quad (14)$$

and

$$\begin{aligned} \beta_u = E[ & ((\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X))) \cdot \theta_{10}(X) \\ & - ((\eta(1, X) - \bar{y}) \cdot p(X) + (\underline{y} - \eta(0, X)) \cdot (1 - p(X))) \cdot \theta_{01}(X)]. \end{aligned} \quad (15)$$

Worst-case analysis is a great starting point as no additional assumptions are imposed on the unobservables. However, the bounds could be too wide to be informative in some cases. In fact, the worst-case bound cover 0 all the time as  $\underline{\beta} \leq 0$  and  $\beta_u \geq 0$ . One could impose additional assumptions on the relationship between the unobservables and obtain tighter bounds. Towards that end, I analyze the monotone treatment response (MTR) assumption of Manski (1997).

**Assumption 1** (MTR Assumption).

$$Y_1 \geq Y_0 \text{ a.s.} \quad (16)$$

Assumption 1 states that everyone benefits from the treatment. Suppose Assumption 1 holds. Then, the random set is such that

$$\mathcal{Y}_1 \times \mathcal{Y}_0 = \begin{cases} \{Y\} \times [\underline{y}, Y] & \text{if } D = 1, \\ [Y, \bar{y}] \times \{Y\} & \text{if } D = 0. \end{cases} \quad (17)$$

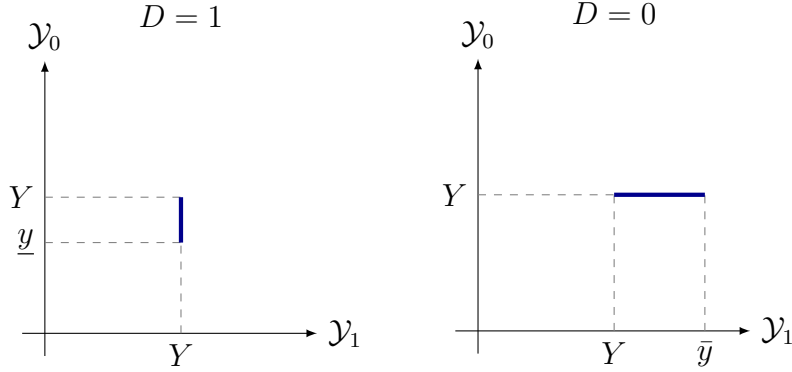


Figure 4: Random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  under MTR Assumption

As shown in Figure 4, depending on the value of  $D$ , the random set in (17) switches its value between two sets, that are smaller than those in (12). The bounds of the welfare gain when the random set is given by (17) are given in the following Corollary to Theorem 1. Notice that the lower bound on conditional average treatment effect  $\underline{\Delta}(X) = -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$  equals 0 when the random set is given by (17). It is shown geometrically in Figure 5. The expected support function of the random set in (17) at direction  $-v^* = (-1, 1)'$  is always 0 as the hyperplane tangent to the random set at direction  $-v^* = (-1, 1)'$  goes through the origin regardless of the value of  $D$ .

**Corollary 2** (MTR). *Suppose Assumption 1 holds. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Then,  $B_I(\delta, \delta^*)$  in (7) is an interval  $[\beta_l, \beta_u]$  where*

$$\beta_l = E\left[-((\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X))) \cdot \theta_{01}(X)\right], \quad (18)$$

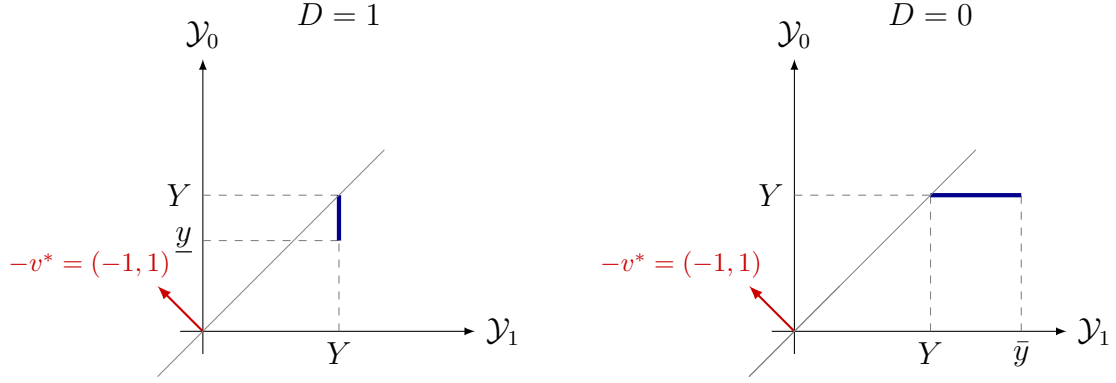


Figure 5: Support function of  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  at direction  $-v^* = (-1, 1)'$  under MTR Assumption

and

$$\beta_u = E\left[\left((\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X))\right) \cdot \theta_{10}(X)\right]. \quad (19)$$

### 3.4 Identification with Instruments

Availability of additional variables, called *instrumental variables*, could help us tighten the bounds on CATE and hence the bounds on the welfare gain. In this subsection, I consider two types of assumptions: (1) mean independence (IV Assumption) and (2) mean monotonicity (MIV Assumption).

#### 3.4.1 Mean independence

**Assumption 2** (IV Assumption). *There exists an instrumental variable  $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$  such that, for  $d \in \{0, 1\}$ , the following mean-independence holds:*

$$E[Y_d | X, Z = z] = E[Y_d | X, Z = z'], \quad (20)$$

for all  $z, z' \in \mathcal{Z}$ .

When data are obtained from a randomized experiment with imperfect compliance, the random assignment can be used as an instrumental variable to identify the effect of the treatment.

Suppose Assumption 2 holds. Since I am imposing an additional restriction on  $(Y_1, Y_0)$ , the

sharp identified region of the welfare gain is given by

$$B_I(\delta, \delta^*) \equiv \{\beta \in \mathbb{R} : \beta = E[E[Y_1 - Y_0|X] \cdot (\delta(X) - \delta^*(X))], (Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0), \\ (Y_1, Y_0) \text{ satisfies Assumption 2}\}. \quad (21)$$

The following lemma corresponds to the Manski's sharp bounds for CATE under mean-independence assumption. [Manski \(1990\)](#) explains it for the more general case of when there are level-set restrictions on the outcome regression.

**Lemma 2 (IV).** *Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  be an integrable random set that is almost surely convex and let  $(Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ . Let  $v_1 = (1, 0)'$  and  $v_0 = (0, 1)'$ . Suppose Assumption 2 holds. Then, we have*

$$\begin{aligned} \sup_{z \in \mathcal{Z}} \{ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} - \inf_{z \in \mathcal{Z}} \{ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} \\ \leq E[Y_1 - Y_0|X] \leq \\ \inf_{z \in \mathcal{Z}} \{ E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} - \sup_{z \in \mathcal{Z}} \{ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} \quad a.s. \end{aligned} \quad (22)$$

Bounds for CATE with instrumental variables involve expected support functions at directions  $v_1 = (1, 0)$  and  $v_0 = (0, 1)$ . The support function of the random set  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  at direction  $v_1 = (1, 0)$  under worst-case is depicted in Figure 6.

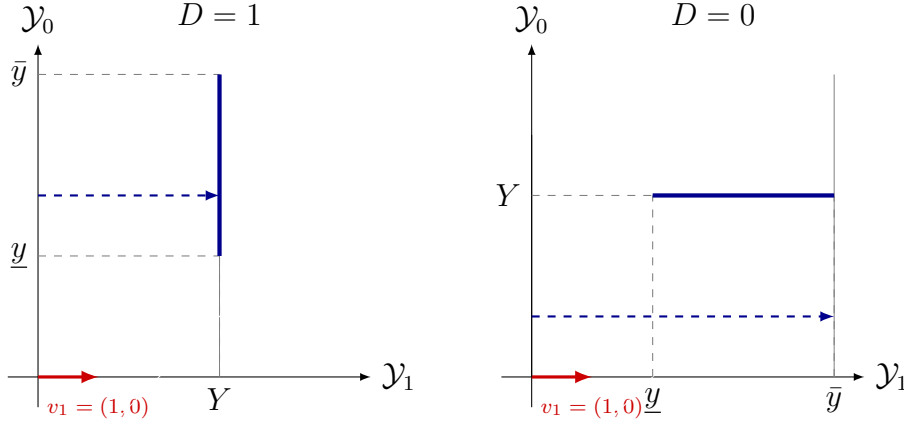


Figure 6: Support function of  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  at direction  $v_1 = (1, 0)$  under worst-case

**Theorem 2 (IV).** Suppose  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  is an integrable random set that is almost surely convex. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Also, let  $v_1 = (1, 0)'$  and  $v_0 = (0, 1)'$ . Then,  $B_I(\delta, \delta^*)$  in (21) is an interval  $[\beta_l, \beta_u]$  where

$$\beta_l = E[\underline{\Delta}(X) \cdot \theta_{10}(X) - \bar{\Delta}(X) \cdot \theta_{01}(X)], \quad (23)$$

and

$$\beta_u = E[\bar{\Delta}(X) \cdot \theta_{10}(X) - \underline{\Delta}(X) \cdot \theta_{01}(X)], \quad (24)$$

where  $\underline{\Delta}(X) \equiv \sup_{z \in \mathcal{Z}} \{ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} - \inf_{z \in \mathcal{Z}} \{ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \}$  and  $\bar{\Delta}(X) \equiv \inf_{z \in \mathcal{Z}} \{ E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} - \sup_{z \in \mathcal{Z}} \{ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \}.$

Identification of the welfare gain with instruments is similar to identification without instruments. The difference lies in the forms of lower and upper bounds on the CATE. Theorem 2 can be combined with different maintained assumptions on the potential outcomes to result in different bounds. Corollary 3 shows the IV bounds under worst-case assumption and Corollary 4 shows the IV bounds under MTR assumption. To ease notation, let  $\eta(d, x, z) \equiv E[Y|D = d, X = x, Z = z]$  for  $d \in \{0, 1\}$  denote the conditional mean treatment responses and  $p(x, z) \equiv P(D = 1|X = x, Z = z)$  denote the propensity score.

**Corollary 3 (IV-worst case).** Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0)$  be a random set in (12). Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Then,  $B_I(\delta, \delta^*)$  in (21) is an interval  $[\beta_l, \beta_u]$  where

$$\begin{aligned} \beta_l = E[ & \left( \sup_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \underline{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \inf_{z \in \mathcal{Z}} \{ \bar{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{10}(X) \\ & - \left( \inf_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \sup_{z \in \mathcal{Z}} \{ \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{01}(X) ], \end{aligned} \quad (25)$$

and

$$\begin{aligned} \beta_u = E[ & \left( \inf_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \sup_{z \in \mathcal{Z}} \{ \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{10}(X) \\ & - \left( \sup_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \underline{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \inf_{z \in \mathcal{Z}} \{ \bar{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{01}(X) ]. \end{aligned} \quad (26)$$



**Corollary 4** (IV-MTR). *Suppose Assumption 1 holds. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Then,  $B_I(\delta, \delta^*)$  in (21) is an interval  $[\beta_l, \beta_u]$  where*

$$\begin{aligned} \beta_l = E \Big[ & - \left( \inf_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \sup_{z \in \mathcal{Z}} \{ \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{01}(X) \Big], \end{aligned} \quad (27)$$

and

$$\begin{aligned} \beta_u = E \Big[ & \left( \inf_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)) \} \right. \\ & \left. - \sup_{z \in \mathcal{Z}} \{ \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \} \right) \cdot \theta_{10}(X) \Big]. \end{aligned} \quad (28)$$

Bounds obtained with instruments are functions of  $\eta(1, x, z)$ ,  $\eta(0, x, z)$  and  $p(x, z)$  and involve taking intersections across values of  $Z$ . If  $Z$  is continuous, this would amount to infinitely many intersections. However, bounds can be simplified in some empirically relevant cases such as the following.

**Assumption 3** (Binary IV with monotonic first-step). *Suppose  $Z \in \{0, 1\}$  is a binary instrumental variable that satisfies Assumption 2. Suppose further that for all  $x \in \mathcal{X}$ ,*

$$p(x, 1) = P(D = 1 | X = x, Z = 1) \geq P(D = 1 | X = x, Z = 0) = p(x, 0). \quad (29)$$

When  $Z \in \{0, 1\}$  is random offer and  $D \in \{0, 1\}$  is program participation, this means that someone who received an offer to participate in the program is more likely to participate in the program than someone who didn't receive an offer.

**Lemma 3.** *Suppose Assumption 3 holds. Then,*

$$1 = \arg \max_{z \in \{0, 1\}} \{ \eta(1, X, z) \cdot p(X, z) + \underline{y} \cdot (1 - p(X, z)) \}, \quad (30)$$

$$0 = \arg \min_{z \in \{0, 1\}} \{ \bar{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \}, \quad (31)$$

$$1 = \arg \min_{z \in \{0, 1\}} \{ \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)) \}, \quad (32)$$

$$0 = \arg \max_{z \in \{0, 1\}} \{ \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)) \}. \quad (33)$$

Under Assumption 3, using Lemma 3, bounds in (25) and (26) are simplified as

$$\begin{aligned}\beta_l = E[ & ((\eta(1, X, 1) \cdot p(X, 1) + \underline{y} \cdot (1 - p(X, 1))) \\ & - (\bar{y} \cdot p(X, 0) + \eta(0, X, 0) \cdot (1 - p(X, 0)))) \cdot \theta_{10}(X) \\ & - ((\eta(1, X, 1) \cdot p(X, 1) + \bar{y} \cdot (1 - p(X, 1))) \\ & - (\underline{y} \cdot p(X, 0) + \eta(0, X, 0) \cdot (1 - p(X, 0)))) \cdot \theta_{01}(X)],\end{aligned}\quad (34)$$

and

$$\begin{aligned}\beta_u = E[ & ((\eta(1, X, 1) \cdot p(X, 1) + \bar{y} \cdot (1 - p(X, 1))) \\ & - (\underline{y} \cdot p(X, 0) + \eta(0, X, 0) \cdot (1 - p(X, 0)))) \cdot \theta_{10}(X) \\ & - ((\eta(1, X, 1) \cdot p(X, 1) + \underline{y} \cdot (1 - p(X, 1))) \\ & - (\bar{y} \cdot p(X, 0) + \eta(0, X, 0) \cdot (1 - p(X, 0)))) \cdot \theta_{01}(X)].\end{aligned}\quad (35)$$

Bounds in (27) and (28) can also be simplified similarly.

### 3.4.2 Mean monotonicity

Next, I consider monotone instrumental variable (MIV) assumption introduced by Manski and Pepper (2000) which weakens Assumption 2 by replacing the equality in (20) by an inequality. An instrumental variable which satisfies this assumption could also help us obtain tighter bounds.

**Assumption 4** (MIV Assumption). *There exists an instrumental variable  $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$  such that, for  $d \in \{0, 1\}$ , the following mean monotonicity holds:*

$$E[Y_d|X, Z = z] \geq E[Y_d|X, Z = z'], \quad (36)$$

for all  $z, z' \in \mathcal{Z}$  such that  $z \geq z'$ .

In the job training program example, the pre-program earnings can be used as an monotone instrumental variable when the outcome variable is post-program earnings.

Suppose Assumption 4 holds. Then, the sharp identified region of the welfare gain is given by

$$\begin{aligned}B_I(\delta, \delta^*) \equiv & \{\beta \in \mathbb{R} : \beta = E[E[Y_1 - Y_0|X] \cdot (\delta(X) - \delta^*(X))], (Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0), \\ & (Y_1, Y_0) \text{ satisfies Assumption 4}\}.\end{aligned}\quad (37)$$

**Lemma 4** (MIV). *Let  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  be an integrable random set that is almost surely convex and let  $(Y_1, Y_0) \in \mathcal{S}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ . Let  $v_1 = (1, 0)'$  and  $v_0 = (0, 1)'$ . Suppose Assumption 4 holds. Then, we have*

$$\begin{aligned} \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_1] \} - \inf_{z_2 \geq z} \{ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_2] \} \right) \\ \leq E[Y_1 - Y_0 | X] \leq \\ \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_2] \} - \sup_{z_1 \leq z} \{ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_1] \} \right) \quad a.s. \end{aligned} \quad (38)$$

**Theorem 3** (MIV). *Suppose  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  is an integrable random set that is almost surely convex. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Also, let  $v_1 = (1, 0)'$  and  $v_0 = (0, 1)'$ . Then,  $B_I(\delta, \delta^*)$  in (37) is an interval  $[\beta_l, \beta_u]$  where*

$$\beta_l = E[\underline{\Delta}(X) \cdot \theta_{10}(X) - \bar{\Delta}(X) \cdot \theta_{01}(X)], \quad (39)$$

and

$$\beta_u = E[\bar{\Delta}(X) \cdot \theta_{10}(X) - \underline{\Delta}(X) \cdot \theta_{01}(X)], \quad (40)$$

where

$$\underline{\Delta}(X) \equiv \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_1] \} - \inf_{z_2 \geq z} \{ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_2] \} \right),$$

and

$$\bar{\Delta}(X) \equiv \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_2] \} - \sup_{z_1 \leq z} \{ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_1] \} \right).$$

**Corollary 5** (MIV-worst case). *Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules.*

Then,  $B_I(\delta, \delta^*)$  in (37) is an interval  $[\beta_l, \beta_u]$  where

$$\begin{aligned} \beta_l = E \Big[ & \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ \eta(1, X, z_1) \cdot p(X, z_1) + \underline{y} \cdot (1 - p(X, z_1)) \} \right. \\ & \left. - \inf_{z_2 \geq z} \{ \bar{y} \cdot p(X, z_2) + \eta(0, X, z_2) \cdot (1 - p(X, z_2)) \} \right) \cdot \theta_{10}(X) \\ & - \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ \eta(1, X, z_2) \cdot p(X, z_2) + \bar{y} \cdot (1 - p(X, z_2)) \} \right. \\ & \left. - \sup_{z_1 \leq z} \{ \underline{y} \cdot p(X, z_1) + \eta(0, X, z_1) \cdot (1 - p(X, z_1)) \} \right) \cdot \theta_{01}(X) \Big], \end{aligned} \quad (41)$$

and

$$\begin{aligned} \beta_u = E \Big[ & \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ \eta(1, X, z_2) \cdot p(X, z_2) + \bar{y} \cdot (1 - p(X, z_2)) \} \right. \\ & \left. - \sup_{z_1 \leq z} \{ \underline{y} \cdot p(X, z_1) + \eta(0, X, z_1) \cdot (1 - p(X, z_1)) \} \right) \cdot \theta_{10}(X) \\ & - \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ \eta(1, X, z_1) \cdot p(X, z_1) + \underline{y} \cdot (1 - p(X, z_1)) \} \right. \\ & \left. - \inf_{z_2 \geq z} \{ \bar{y} \cdot p(X, z_2) + \eta(0, X, z_2) \cdot (1 - p(X, z_2)) \} \right) \cdot \theta_{01}(X) \Big]. \end{aligned} \quad (42)$$

**Corollary 6** (MIV-MTR). *Suppose Assumption 1 holds. Let  $\delta : \mathcal{X} \rightarrow \{0, 1\}$  and  $\delta^* : \mathcal{X} \rightarrow \{0, 1\}$  be treatment rules. Then,  $B_I(\delta, \delta^*)$  in (37) is an interval  $[\beta_l, \beta_u]$  where*

$$\begin{aligned} \beta_l = E \Big[ & \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ E[Y|X, Z = z_1] \} - \inf_{z_2 \geq z} \{ E[Y|X, Z = z_2] \} \right) \cdot \theta_{10}(X) \\ & - \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ \eta(1, X, z_2) \cdot p(X, z_2) + \bar{y} \cdot (1 - p(X, z_2)) \} \right. \\ & \left. - \sup_{z_1 \leq z} \{ \underline{y} \cdot p(X, z_1) + \eta(0, X, z_1) \cdot (1 - p(X, z_1)) \} \right) \cdot \theta_{01}(X) \Big], \end{aligned} \quad (43)$$

and

$$\begin{aligned} \beta_u = E \Big[ & \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ \eta(1, X, z_2) \cdot p(X, z_2) + \bar{y} \cdot (1 - p(X, z_2)) \} \right. \\ & \left. - \sup_{z_1 \leq z} \{ \underline{y} \cdot p(X, z_1) + \eta(0, X, z_1) \cdot (1 - p(X, z_1)) \} \right) \cdot \theta_{10}(X) \\ & - \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ E[Y|X, Z = z_1] \} - \inf_{z_2 \geq z} \{ E[Y|X, Z = z_2] \} \right) \cdot \theta_{01}(X) \Big]. \end{aligned} \quad (44)$$

Table 1 summarizes the forms of lower and upper bounds on CATE under different sets of assumptions.

Table 1

Assumptions	$\underline{\Delta}(X)$	$\bar{\Delta}(X)$
worst-case	$(\eta(1, X) - \bar{y}) \cdot p(X) \\ + (\underline{y} - \eta(0, X)) \cdot (1 - p(X))$	$(\eta(1, X) - \underline{y}) \cdot p(X) \\ + (\bar{y} - \eta(0, X)) \cdot (1 - p(X))$
MTR	0	same as worst-case
IV-worst-case	$\sup_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) \\ + \underline{y} \cdot (1 - p(X, z)) \} \\ - \inf_{z \in \mathcal{Z}} \{ \bar{y} \cdot p(X, z) \\ + \eta(0, X, z) \cdot (1 - p(X, z)) \}$	$\inf_{z \in \mathcal{Z}} \{ \eta(1, X, z) \cdot p(X, z) \\ + \bar{y} \cdot (1 - p(X, z)) \} \\ - \sup_{z \in \mathcal{Z}} \{ \underline{y} \cdot p(X, z) \\ + \eta(0, X, z) \cdot (1 - p(X, z)) \}$
IV-MTR	0	same as IV-worst-case
MIV-worst-case	$\sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ \eta(1, X, z_1) \cdot p(X, z_1) \\ + \underline{y} \cdot (1 - p(X, z_1)) \} \right. \\ \left. - \inf_{z_2 \geq z} \{ \bar{y} \cdot p(X, z_2) \\ + \eta(0, X, z_2) \cdot (1 - p(X, z_2)) \} \right)$	$\sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \{ \eta(1, X, z_2) \cdot p(X, z_2) \\ + \bar{y} \cdot (1 - p(X, z_2)) \} \right. \\ \left. - \sup_{z_1 \leq z} \{ \underline{y} \cdot p(X, z_1) \\ + \eta(0, X, z_1) \cdot (1 - p(X, z_1)) \} \right)$
MIV-MTR	$\sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \{ E[Y X, Z = z_1] \} \right. \\ \left. - \inf_{z_2 \geq z} \{ E[Y X, Z = z_2] \} \right)$	same as MIV-worst-case

This table reports the form of  $\underline{\Delta}(X)$  and  $\bar{\Delta}(X)$  under different assumptions.

## 4 Estimation and Inference

The bounds developed in Section 3 are functions of conditional mean treatment responses  $\eta(1, x)$  and  $\eta(0, x)$ , and propensity score  $p(x)$  in the absence of instruments. The bounds with instruments are functions of conditional mean treatment responses  $\eta(1, x, z)$  and  $\eta(0, x, z)$ , and propensity score  $p(x, z)$ . Let  $F$  be the joint distribution of  $W = (Y, D, X, Z)$  and suppose we have a size  $n$  random sample  $\{w_i\}_{i=1}^n$  from  $W$ .

If the conditioning variables  $X$  and  $Z$  are discrete and take finitely many values, conditional mean treatment responses and propensity scores can be estimated by the corresponding empirical means. If there is a continuous component, conditional mean treatment responses and propensity scores can be estimated using nonparametric regression methods. I start with bounds that do not rely on instruments. Let  $\hat{\eta}(1, x)$ ,  $\hat{\eta}(0, x)$ , and  $\hat{p}(x)$  be those estimated values. A natural sample analog estimator for the lower bound under the worst-case in (14) can be constructed by first plugging these estimated values into (14) and then by taking average over  $i$  as follows:

$$\begin{aligned} \hat{\beta}_l = \frac{1}{n} \sum_{i=1}^n & [((\hat{\eta}(1, x_i) - \bar{y}) \cdot \hat{p}(x_i) + (\underline{y} - \hat{\eta}(0, x_i)) \cdot (1 - \hat{p}(x_i))) \cdot \theta_{10}(x_i) \\ & - ((\hat{\eta}(1, x_i) - \underline{y}) \cdot \hat{p}(x_i) + (\bar{y} - \hat{\eta}(0, x_i)) \cdot (1 - \hat{p}(x_i))) \cdot \theta_{01}(x_i)]. \end{aligned} \quad (45)$$

In this estimation problem,  $\eta(1, x)$ ,  $\eta(0, x)$ , and  $p(x)$  are nuisance parameters that need to be estimated nonparametrically. In what follows, I collect these possibly infinite-dimensional nuisance parameters and denote it as follows:<sup>4</sup>

$$\gamma = (\eta(1, \cdot), \eta(0, \cdot), p(\cdot)). \quad (46)$$

Estimation of these parameters can affect the sampling distribution of  $\hat{\beta}_l$  in a complicated manner. To mitigate the effect of this first-step nonparametric estimation, one could use an orthogonalized moment condition, which I describe below, to estimate  $\beta_l$ .

Let  $\beta_*$  denote either the lower bound or the upper bound, i.e.,  $\beta_* \in \{\beta_l, \beta_u\}$ . I write my estimator as a generalized method of moments (GMM) estimator in which the true value  $\beta_{*,0}$  of  $\beta_*$  satisfies a single moment restriction

$$E[m(w_i, \beta_{*,0}, \gamma_0)] = 0, \quad (47)$$

---

<sup>4</sup>I use  $\eta(1, \cdot)$ ,  $\eta(0, \cdot)$ , and  $p(\cdot)$  instead of  $\eta(1, x)$ ,  $\eta(0, x)$ , and  $p(x)$  to highlight the fact that they are functions.

where

$$m(w, \beta_l, \gamma) = \underline{\Delta}(\gamma) \cdot \theta_{10}(x) - \bar{\Delta}(\gamma) \cdot \theta_{01}(x) - \beta_l, \quad (48)$$

and

$$m(w, \beta_u, \gamma) = \bar{\Delta}(\gamma) \cdot \theta_{10}(x) - \underline{\Delta}(\gamma) \cdot \theta_{01}(x) - \beta_u. \quad (49)$$

$\underline{\Delta}(\gamma)$  and  $\bar{\Delta}(\gamma)$  denote the lower and upper bound on CATE respectively and are functions of the nuisance parameters  $\gamma$ .

We would like our moment function to have an *orthogonality* property so that the estimation of parameter of interest would be first-order insensitive to nonparametric estimation errors in the nuisance parameter. This allows for the use of various nonparametric estimators of these parameters including high-dimensional machine learning estimators. I construct such moment function by adding influence function adjustment term for first step estimation  $\phi(w, \beta_*, \gamma)$  to the original moment function  $m(w, \beta_*, \gamma)$  as in CEINR. Let the *orthogonalized* moment function be denoted by

$$\psi(w, \beta_*, \gamma) = m(w, \beta_*, \gamma) + \phi(w, \beta_*, \gamma). \quad (50)$$

Let  $F_\tau = (1 - \tau)F_0 + \tau G$  for  $\tau \in [0, 1]$ , where  $F_0$  is the true distribution of  $W$  and  $G$  is some alternative distribution. Then, we say that the moment condition satisfies the Neyman orthogonality condition or is locally robust if

$$\left. \frac{d}{d\tau} E[\psi(w_i, \beta_{*,0}, \gamma(F_\tau))] \right|_{\tau=0} = 0. \quad (51)$$

The orthogonality has been used in semiparametric problems by [Newey \(1990, 1994\)](#), [Andrews \(1994\)](#), [Robins and Rotnitzky \(1995\)](#), among others. More recently, in a high-dimensional setting, it has been used by [Belloni, Chen, Chernozhukov, and Hansen \(2012\)](#), [Belloni, Chernozhukov, and Hansen \(2014\)](#), [Farrell \(2015\)](#), [Belloni, Chernozhukov, Fernández-Val, and Hansen \(2017\)](#), [Athey, Imbens, and Wager \(2018\)](#), and [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins \(2018\)](#), among others. Recently, [Sasaki and Ura \(2018\)](#) proposed using orthogonalized moments for the estimation and inference of a parameter called policy relevant treatment effect (PRTE) whose explanation can be found in [Heckman and Vytlacil \(2007\)](#). Much like our problem, the estimation of the PRTE involves estimation of multiple nuisance parameters.



## 4.1 Influence function calculation

In this subsection, I show how I derive the adjustment term  $\phi(w, \beta_l, \gamma)$  for the lower bound under the worst-case assumption. This illustrates how I derive the adjustment term for the cases in which  $\underline{\Delta}(\gamma)$  and  $\bar{\Delta}(\gamma)$  are differentiable with respect to  $\gamma$ , i.e., cases in which we do not have instrumental variables. Additional assumptions need to be imposed for the cases where  $\underline{\Delta}(\gamma)$  and  $\bar{\Delta}(\gamma)$  are non-differentiable with respect to  $\gamma$ .

Under the worst-case assumption, the original moment function for lower bound takes the following form:

$$\begin{aligned} m(w, \beta_l, \gamma) = & ((\eta(1, x) - \bar{y}) \cdot p(x) + (\underline{y} - \eta(0, x)) \cdot (1 - p(x))) \cdot \theta_{10}(x) \\ & - ((\eta(1, x) - \underline{y}) \cdot p(x) + (\bar{y} - \eta(0, x)) \cdot (1 - p(x))) \cdot \theta_{01}(x) - \beta_l. \end{aligned} \quad (52)$$

**Assumption 5.**  $\eta(1, x), \eta(0, x)$ , and  $p(x)$  are continuous at every  $x$ .

**Lemma 5.** If Assumption 5 is satisfied then the influence function of  $E[m(w, \beta_{l,0}, \gamma(F))]$  is  $\phi(w, \beta_{l,0}, \gamma_0)$  which is given by

$$\phi(w, \beta_{l,0}, \gamma_0) = \phi_1 + \phi_2, \quad (53)$$

where

$$\begin{aligned} \phi_1 &= (\theta_{10}(x) - \theta_{01}(x)) \cdot (\eta_0(1, x) + \eta_0(0, x) - (\underline{y} + \bar{y})) \cdot (d - p_0(x)), \\ \phi_2 &= (\theta_{10}(x) - \theta_{01}(x)) \cdot [y - \eta_0(1, x)]^d \cdot [-(y - \eta_0(0, x))]^{1-d}. \end{aligned} \quad (54)$$

Note that we have  $E[\phi(w, \beta_{l,0}, \gamma_0)] = 0$  so that the orthogonalized moment condition  $\psi(w, \beta_l, \gamma)$  still identifies our parameter of interest with  $E[\psi(w, \beta_{l,0}, \gamma_0)] = 0$ . The adjustment term consists of two terms. While term  $\phi_1$  represents the effect of local perturbations of the distribution of  $D|X$  on the moment, term  $\phi_2$  represents the effect of local perturbations of the distribution of  $Y|D, X$  on the moment.

## 4.2 GMM estimator and its asymptotic variance

Following CEINR, I use cross-fitting, a version of sample splitting, in the construction of sample moments. Cross-fitting works as follows. Let  $K > 1$  be a number of folds. Partitioning the set of observation indices  $\{1, 2, \dots, n\}$  into  $K$  groups  $\mathcal{I}_k, k = 1, \dots, K$ , let  $\hat{\gamma}_k$  be the first step estimates constructed from all observations not in  $\mathcal{I}_k$ . Then,  $\hat{\beta}_*$  can be obtained as a solution

to

$$\frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \psi(w_i, \hat{\beta}_*, \hat{\gamma}_k) = 0. \quad (55)$$

**Assumption 6.** For each  $k = 1, \dots, K$ , (i)  $\int \|\psi(w, \beta_{*,0}, \hat{\gamma}_k) - \psi(w, \beta_{*,0}, \gamma_0)\|^2 F_0(dw) \xrightarrow{p} 0$ , (ii)  $\|\int \psi(w, \beta_{*,0}, \hat{\gamma}_k) F_0(dw)\| \leq C \|\hat{\gamma}_k - \gamma_0\|^2$  for  $C > 0$ , (iii)  $\|\hat{\gamma}_k - \gamma_0\| = o_p(n^{-1/4})$  (iv) there is  $\zeta > 0$  and  $d(w_i)$  with  $E[d(w_i)^2] < \infty$  such that for  $\|\beta_* - \beta_{*,0}\|$  and  $\|\gamma - \gamma_0\|$  small enough

$$|\psi(w_i, \beta_*, \gamma) - \psi(w_i, \beta_{*,0}, \gamma_0)| \leq d(w_i)(\|\beta_* - \beta_{*,0}\|^\zeta + \|\gamma - \gamma_0\|^\zeta).$$

**Theorem 4.** Suppose that  $\{w_i\}_{i=1}^n$  are i.i.d., Assumption 6 (i), (ii), and (iii) are satisfied,  $\hat{\beta}_* \xrightarrow{p} \beta_{*,0}$ , and  $\Omega_* \equiv E[\psi(w_i, \beta_{*,0}, \gamma_0)^2] < \infty$ . Then

$$\sqrt{n}(\hat{\beta}_* - \beta_{*,0}) \xrightarrow{d} \mathcal{N}(0, \Omega_*). \quad (56)$$

Moreover, if Assumption 6 (iv) is also satisfied, a consistent estimator for the asymptotic variance can be constructed as

$$\hat{\Omega}_* = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \psi(w_i, \hat{\beta}_*, \hat{\gamma}_k)^2. \quad (57)$$

**Corollary 7** (Locally robust estimator of the lower bound under the worst-case and a consistent estimator of its asymptotic variance). A locally robust estimator  $\hat{\beta}_l$  of the lower bound under the worst-case takes the form

$$\begin{aligned} \hat{\beta}_l = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} & \left[ ((\hat{\eta}_k(1, x_i) - \bar{y}) \cdot \hat{p}_k(x_i) + (\underline{y} - \hat{\eta}_k(0, x_i)) \cdot (1 - \hat{p}_k(x_i))) \cdot \theta_{10}(x_i) \right. \\ & - ((\hat{\eta}_k(1, x_i) - \underline{y}) \cdot \hat{p}_k(x_i) + (\bar{y} - \hat{\eta}_k(0, x_i)) \cdot (1 - \hat{p}_k(x_i))) \cdot \theta_{01}(x_i) \\ & + (\theta_{10}(x_i) - \theta_{01}(x_i)) \cdot (\hat{\eta}_k(1, x_i) + \hat{\eta}_k(0, x_i) - (\underline{y} + \bar{y})) \cdot (d_i - \hat{p}_k(x_i)) \\ & \left. + (\theta_{10}(x_i) - \theta_{01}(x_i)) \cdot [y_i - \hat{\eta}_k(1, x_i)]^{d_i} \cdot [-(y_i - \hat{\eta}_k(0, x_i))]^{1-d_i} \right]. \end{aligned} \quad (58)$$

Moreover, a consistent estimator of its asymptotic variance takes the form

$$\begin{aligned}
\hat{\Omega}_l &= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \psi(w_i, \hat{\beta}_l, \hat{\gamma}_k)^2 \\
&= \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \left[ \left( (\hat{\eta}_k(1, x_i) - \bar{y}) \cdot \hat{p}_k(x_i) + (\underline{y} - \hat{\eta}_k(0, x_i)) \cdot (1 - \hat{p}_k(x_i)) \right) \cdot \theta_{10}(x_i) \right. \\
&\quad - \left( (\hat{\eta}_k(1, x_i) - \underline{y}) \cdot \hat{p}_k(x_i) + (\bar{y} - \hat{\eta}_k(0, x_i)) \cdot (1 - \hat{p}_k(x_i)) \right) \cdot \theta_{01}(x_i) - \hat{\beta}_l \\
&\quad + (\theta_{10}(x_i) - \theta_{01}(x_i)) \cdot (\hat{\eta}_k(1, x_i) + \hat{\eta}_k(0, x_i) - (\underline{y} + \bar{y})) \cdot (d_i - \hat{p}_k(x_i)) \\
&\quad \left. + (\theta_{10}(x_i) - \theta_{01}(x_i)) \cdot [y_i - \hat{\eta}_k(1, x_i)]^{d_i} \cdot [-(y_i - \hat{\eta}_k(0, x_i))]^{1-d_i} \right]^2.
\end{aligned} \tag{59}$$

Given locally robust estimators  $\hat{\beta}_l$  and  $\hat{\beta}_u$  of the lower and upper bound  $\beta_l$  and  $\beta_u$ , and consistent estimators  $\hat{\Omega}_l$  and  $\hat{\Omega}_u$  of their asymptotic variance  $\Omega_l$  and  $\Omega_u$ , we can construct the  $100 \cdot \alpha\%$  confidence interval for the lower bound  $\beta_l$  and upper bound  $\beta_u$  as

$$CI_{\alpha}^{\beta_l} = [\hat{\beta}_l - C_{\alpha} \cdot (\hat{\Omega}_l/n)^{1/2}, \hat{\beta}_l + C_{\alpha} \cdot (\hat{\Omega}_l/n)^{1/2}], \tag{60}$$

and

$$CI_{\alpha}^{\beta_u} = [\hat{\beta}_u - C_{\alpha} \cdot (\hat{\Omega}_u/n)^{1/2}, \hat{\beta}_u + C_{\alpha} \cdot (\hat{\Omega}_u/n)^{1/2}], \tag{61}$$

where  $C_{\alpha}$  satisfies

$$\Phi(C_{\alpha}) - \Phi(-C_{\alpha}) = \alpha. \tag{62}$$

In other words,  $C_{\alpha}$  is the value that satisfies  $\Phi(C_{\alpha}) = (\alpha + 1)/2$ , i.e, the  $(\alpha + 1)/2$  quantile of the standard normal distribution. For example, when  $\alpha = 0.95$ ,  $C_{\alpha}$  is 1.96.

### 4.3 Bounds with instruments

When there are additional instrumental variables,  $\underline{\Delta}(\gamma)$  and  $\bar{\Delta}(\gamma)$  in (48) and (49) are non-differentiable with respect to  $\gamma$  as they involve sup and inf operators. However, under additional monotonicity assumption, the bounds can be simplified. In this section, I derive the influence function for the IV-worst-case lower bound under the monotonicity assumption. Under mono-

tonicity, the moment condition for the IV-worst-case lower bound is

$$\begin{aligned}
m(w, \beta_l, \gamma) = & (\eta(1, x, 1) \cdot p(x, 1) + \underline{y} \cdot (1 - p(x, 1)) \\
& - \bar{y} \cdot p(x, 0) - \eta(0, x, 0) \cdot (1 - p(x, 0))) \cdot \theta_{10}(x) \\
& - (\eta(1, x, 1) \cdot p(x, 1) + \bar{y} \cdot (1 - p(x, 1)) \\
& - \underline{y} \cdot p(x, 0) - \eta(0, x, 0) \cdot (1 - p(x, 0))) \cdot \theta_{01}(x) - \beta_l.
\end{aligned} \tag{63}$$

**Lemma 6.** *If Assumption 5 is satisfied then the influence function of  $E[m(w, \beta_{l,0}, \gamma(F))]$  is  $\phi(w, \beta_{l,0}, \gamma_0)$  which is given by*

$$\phi(w, \beta_{l,0}, \gamma_0) = \phi_1 + \phi_2, \tag{64}$$

where

$$\begin{aligned}
\phi_1 = & [((\eta_0(1, x, 1) - \underline{y}) \cdot \theta_{10}(x) - (\eta_0(1, x, 1) - \bar{y}) \cdot \theta_{01}(x)) \cdot (d - p_0(x, 1)))^z \\
& \cdot [((\eta_0(0, x, 0) - \bar{y}) \cdot \theta_{10}(x) - (\eta_0(0, x, 0) - \underline{y}) \cdot \theta_{01}(x)) \cdot (d - p_0(x, 0))]^{1-z} \\
\phi_2 = & (\theta_{10}(x) - \theta_{01}(x)) \cdot (\mathbb{1}\{d = 1, z = 1\} \cdot (y - \eta_0(1, x, 1)) \\
& + \mathbb{1}\{d = 0, z = 0\} \cdot (-(y - \eta_0(0, x, 0))))).
\end{aligned} \tag{65}$$

Notice again that we have  $E[\phi(w, \beta_{l,0}, \gamma_0)] = 0$  so that the orthogonalized moment condition  $\psi(w, \beta_l, \gamma)$  still identifies our parameter of interest with  $E[\psi(w, \beta_{l,0}, \gamma_0)] = 0$ . The adjustment term again consists of two terms. In this case, while term  $\phi_1$  represents the effect of local perturbations of the distribution of  $D|X, Z$  on the moment, term  $\phi_2$  represents the effect of local perturbations of the distribution of  $Y|D, X, Z$  on the moment.

## 5 Empirical Application

In this section, I illustrate my analysis using experimental data from the National Job Training Partnership Act (JTPA) Study which was commissioned by the U.S. Department of Labor in 1986. The goal of this randomized experiment was to measure the benefits and costs of training programs funded under the JTPA of 1982. Applicants who were randomly assigned to a treatment group were allowed access to the program for 18 months while the ones assigned to a control group were excluded from receiving JTPA services in that period. The original evaluation of the program is based on data of 15,981 applicants. More detailed information about the experiment and program impact estimates can be found in [Bloom, Orr, Bell, Cave, Doolittle, Lin, and Bos \(1997\)](#).

I follow [Kitagawa and Tetenov \(2018\)](#) and focus on adult applicants with available data on 30-month earnings after the random assignment, years of education, and pre-program earnings.<sup>5</sup> Table 2 shows the summary statistics of this sample. The sample consists of 9223 observations, of which 6133 (roughly 2/3) were assigned to the treatment group, and 3090 (roughly 1/3) were assigned to the control group. The means and standard deviations of program participation, 30-month earnings, years of education, and pre-program earnings are given for the entire sample, the treatment group subsample, and the control group subsample.

Treatment variable is the job training program participation and equals 1 for individuals who actually participated in the program. Only 65% of those who got assigned to the treatment group actually participated in the training program. I look at the joint distribution of assigned and realized treatment status in Table 3 to further investigate the compliance issue. Outcome variable is 30-month earnings and is on average \$16,093 and ranges from \$0 to \$155,760 with median earnings \$11,187. In the analysis below, based on this range, I set  $\underline{y} = \$0$  and  $\bar{y} = \$160,000$ . Treatment group assignees earned \$16,487 on average while control group assignees earned \$15,311. The \$1,176 difference between these two group averages is an estimate of the JTPA impact on earnings from an intention-to-treat perspective. Pretreatment covariates I consider are years of education and pre-program earnings. Years of education are on average 11.61 years and range from 7 to 18 years with median 12 years. Pre-program earnings are on average \$3,232 and range from \$0 to \$63,000 with median earnings \$1,600. Not surprisingly, both variables are roughly balanced by assignment status due to random assignment and large samples involved.

---

<sup>5</sup> I downloaded the dataset that [Kitagawa and Tetenov \(2018\)](#) used in their analysis from <https://www.econometricsociety.org/content/supplement-who-should-be-treated-empirical-welfare-maximization-methods-treatment-choice>. I supplemented this dataset with that of [Abadie, Angrist, and Imbens \(2002\)](#), which I downloaded from <https://economics.mit.edu/faculty/angrist/data1/data/abangim02>, to obtain a variable that indicates program participation.

Although the offer of treatment was randomly assigned, the compliance was not perfect. Table 3 shows the joint distribution of assigned and realized treatment. Assigned treatment equals 1 for individuals who got offered the training program and realized treatment equals 1 for individuals who actually participated in the training. As can be seen from this table, the realized treatment is not equal to assigned treatment for roughly 23% of the applicants. Therefore, the program participation is self-selected and likely to be correlated with potential outcomes. Since the assumption of unconfoundedness fails to hold in this case, the treatment effects are not point identified. Although the random offer can be used as a treatment variable to point identify the intention-to-treat effect as in Kitagawa and Tetenov (2018), the actual program participation should be used to identify the treatment effect itself.

Table 2: Summary statistics

	Entire sample	Assigned to treatment	Assigned to control
<i>Treatment</i>			
Job training	0.44 (0.50)	0.65 (0.48)	0.01 (0.12)
<i>Outcome variable</i>			
30-month earnings	16,093 (17,071)	16,487 (17,391)	15,311 (16,392)
<i>Pretreatment covariates</i>			
Years of education	11.61 (1.87)	11.63 (1.87)	11.58 (1.88)
Pre-program earnings	3,232 (4,264)	3,205 (4,279)	3,287 (4,234)
Number of observations	9223	6133	3090

This table reports the means and standard deviations (in brackets) of variables in our sample. Treatment variable is job training program participation and equals 1 for individuals who actually participated in the program. The outcome variable is 30-month earnings after the random assignment. Pretreatment covariates are years of education and pre-program annual earnings. The earnings are in US Dollars.

**Example 1.** Applicants were eligible for training if they faced a certain barriers to employment. This included being a high school dropout. Suppose the benchmark policy is to treat everyone

Table 3: The joint distribution of assigned and realized treatment

Realized treatment	Assigned treatment		Total
	1	0	
1	4015	43	4058
0	2118	3047	5165
Total	6133	3090	9223

This table reports the joint distribution of assigned and realized treatment in our sample. Assigned treatment equals 1 for individuals who got offered job training and realized treatment equals 1 for individuals who actually participated in the training. It shows the compliance issue in our sample.

with less than high school education, i.e., people who have less than or equal to 11 years of education. Now, consider implementing a new policy in which we include people with high school degree. In other words, let

$$\delta^* = \mathbb{1}\{\text{education} \leq 11\}, \quad (66)$$

$$\delta = \mathbb{1}\{\text{education} \leq 12\}. \quad (67)$$

The estimates of lower and upper bounds on the welfare gain from this new policy under various assumptions and different instrumental variables are summarized in Table 4. In this example, a random offer is used as an instrumental variable and pre-program earnings is used as a monotone instrumental variable. For the first-step estimation, I use cross-fitting with  $K = 2$  and estimate  $\hat{\eta}(1, x)$ ,  $\hat{\eta}(0, x)$  and  $\hat{p}(x)$  by empirical means. Those empirical means out of whole sample are depicted in Figure 7 and 8 in the Appendix. Empirical means and distributions when years of education is used as  $X$  and random offer is used as  $Z$  are summarized in Table 7 in the Appendix.

As can be seen from Table 4, the worst-case bounds cover 0, as I explained earlier. Although we cannot rank which policy is better, we quantify the no-assumption scenario as a welfare loss of \$31,423 and a welfare gain of \$36,928. Under the MTR assumption, the lower bound is 0. That is because the MTR assumption states that everyone benefits from the treatment, and under the new policy, we are expanding the treated population. The upper bound under MTR is the same as the upper bound under the worst-case. When we use a random offer as an instrumental variable, the bounds are tighter than the worst-case bounds and still cover 0. However, when we use pre-program earnings as a monotone instrumental variable, the bounds do not cover 0, and it is even tighter if we impose an additional MTR assumption. Therefore, if the researcher is comfortable with the validity of the MIV assumption, she can conclude



that implementing the new policy is guaranteed to improve welfare and that improvement is between \$3,569 and \$36,616.

Table 4: Welfare gains in Example 1

Assumptions	lower bound	upper bound
worst-case	-31,423 (-32,564, -30,282)	36,928 (35,699, 38,158)
MTR	0	36,928 (35699, 38158)
IV-worst-case	-2,486 (-2,774, -2,198)	20,787 (19,881, 21694)
IV-MTR	0	20,787 (19,881, 21694)
MIV-worst-case	3,569	36,616
MIV-MTR	7,167	36,616

This table reports the estimated welfare gains and their 95% confidence intervals (in brackets) in Example 1 under various assumptions. The welfare is in terms of 30-month earnings in US Dollars.

**Example 2.** One class of treatment rules that [Kitagawa and Tetenov \(2018\)](#) considered is a class of quadrant treatment rules:

$$\mathcal{G} = \{ \{x : s_1(\text{education} - t_1) > 0 \text{ and } s_2(\text{pre-program earnings} - t_2) > 0\}, \quad (68)$$

$$s_1, s_2 \in \{-1, 1\}, t_1, t_2 \in \mathbb{R} \}.$$

One's education level and pre-program earnings have to be above or below some specific thresholds to be assigned to treatment according to this treatment rule. Within this class of treatment rules, the empirical welfare maximizing treatment rule that [Kitagawa and Tetenov \(2018\)](#) calculates is  $\mathbb{1}\{\text{education} \leq 15, \text{prior earnings} \leq \$19,670\}$ . Let this policy be the benchmark policy and consider implementing another policy that lowers the education threshold to be 12. In fact, that policy is another empirical welfare maximizing policy that takes into account the treatment assignment cost which is \$774 per assignee. I calculate the welfare difference

between these two policies. In other words, let

$$\delta^* = \mathbb{1}\{\text{education} \leq 15, \text{pre-program earnings} \leq \$19,670\}, \quad (69)$$

$$\delta = \mathbb{1}\{\text{education} \leq 12, \text{pre-program earnings} \leq \$19,670\}. \quad (70)$$

The estimation results are summarized in Table 5. In this example, a random offer is used as an instrumental variable. For the first-step estimation, I use cross-fitting with  $K = 2$  and estimate  $\hat{\eta}(1, x)$  and  $\hat{\eta}(0, x)$  by polynomial regression of degree 2 and  $\hat{p}(x)$  by logistic regression with polynomial of degree 2. Those estimated conditional mean treatment responses and propensity score out of whole sample are depicted in Figure 10 and 11 in the Appendix.

As can be seen from Table 5, again, the worst-case bounds cover 0. However, we quantify the no-assumption scenario as a welfare loss of \$13,435 and a welfare gain of \$11,633. Under the MTR assumption, the upper bound is 0. That is because the MTR assumption states that everyone benefits from the treatment, and under the new policy, we are shrinking the treated population. The lower bound under MTR is the same as the lower bound under the worst-case. When we use a random offer as an instrumental variable, the bounds are tighter and still cover 0 as well. Using IV assumption alone, which is a credible assumption since the offer was randomly assigned in the experiment, we quantify the difference as a welfare loss of \$7,336 and a welfare gain of \$1,035. In this case, the researcher cannot be sure whether implementing the new policy is guaranteed to worsen or improve welfare. However, if she decides that the welfare gain being at most \$1,035 is not high enough, she can go ahead with the first policy.

Table 5: Welfare gains in Example 2

Assumptions	lower bound	upper bound
worst-case	-13,435 (-14,361, -12,510)	11,633 (10,871, 12,394)
MTR	-13,435 (-14,361, -12,510)	0
IV-worst-case	-7,336 (-7,911, -6,763)	1,035 (862, 1,208)
IV-MTR	-7,336 (-7,911, -6,763)	0

This table reports the estimated welfare gains and their 95% confidence intervals (in brackets) in Example 2 under various assumptions. The welfare is in terms of 30-month earnings in US Dollars.

## 6 Simulation Study

Mimicking the empirical application, I consider the following data generating process. Let  $X$  be a discrete random variable with values  $\{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}$  and probability mass function  $\{0.01, 0.06, 0.07, 0.11, 0.13, 0.43, 0.07, 0.06, 0.02, 0.02, 0.01, 0.01\}$ . Conditional on  $X = x$ , let

$$Z|X = x \sim \text{Bernoulli}(2/3), \quad (71)$$

$$U|X = x, Z = z \sim \text{Unif}[0, 1] \text{ for } z \in \{0, 1\}, \quad (72)$$

$$D = \mathbb{1}\{p(X, Z) \geq U\}, \quad (73)$$

$$Y_1|X = x, Z = z, U = u \sim \text{Lognormal}\left(\log \frac{m_1^2(x, u)}{\sqrt{\sigma_1^2 + m_1^2(x, u)}}, \sqrt{\log\left(\frac{\sigma_1^2}{m_1^2(x, u)} + 1\right)}\right), \quad (74)$$

$$Y_0|X = x, Z = z, U = u \sim \text{Lognormal}\left(\log \frac{m_0^2(x, u)}{\sqrt{\sigma_0^2 + m_0^2(x, u)}}, \sqrt{\log\left(\frac{\sigma_0^2}{m_0^2(x, u)} + 1\right)}\right), \quad (75)$$

where

$$p(x, z) = \frac{1}{1 + e^{-(4.89 + 0.05 \cdot x + 5 \cdot z)}}, \quad (76)$$

$$m_1(x, u) = E[Y_1|X = x, Z = z, U = u] = 5591 + 1027 \cdot x + 2000 \cdot u, \quad (77)$$

$$m_0(x, u) = E[Y_0|X = x, Z = z, U = u] = -1127 + 1389 \cdot x + 1000 \cdot u, \quad (78)$$

$$\sigma_1^2 = \text{Var}[Y_1|X = x, Z = z, U = u] = 11000^2, \quad (79)$$

$$\sigma_0^2 = \text{Var}[Y_0|X = x, Z = z, U = u] = 11000^2. \quad (80)$$

In this specification,  $X$  corresponds to years of education and takes values from 7 to 18.  $Z$  corresponds to random offer and follows  $\text{Bernoulli}(2/3)$  to reflect the fact that probability of being randomly assigned to the treatment group is  $2/3$  irrespective of applicants' years of education.  $D$  corresponds to program participation and equals 1 whenever  $p(x, z)$  exceeds the value of  $U$  which is uniformly distributed on  $[0, 1]$ .  $Y_1$  and  $Y_0$  are potential outcomes and observed outcome  $Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$  corresponds to 30-month post-program earnings. For  $d \in \{0, 1\}$ ,  $Y_d$  conditional on  $X$ ,  $Z$ , and  $U$  follows a lognormal distribution whose mean is  $m_d(x, u)$  and variance is  $\sigma_d^2$ . Under this structure, we have

$$E[Y_d|X, Z] = E[Y_d|X] \text{ for } d \in \{0, 1\}. \quad (81)$$

As in Example 1 in Section 5, consider the following pair of policies:

$$\delta^*(x) = 1\{x \leq 11\} \text{ and } \delta(x) = 1\{x \leq 12\}. \quad (82)$$

Policy  $\delta^*$  corresponds to treating everyone who has less than or equal to 11 years of education, and policy  $\delta$  corresponds to treating everyone who has less than or equal to 12 years of education. Then, the population welfare gain is 1,236. The population worst-case bounds are (-31,191, 37,608) and IV-worst-case bounds are (-2,380, 21,227). As in Section 5, I set  $\underline{y} = 0$  and  $\bar{y} = 160,000$  to calculate the bounds. More details on the calculation of these population quantities can be found in Appendix D.

I focus on worst-case lower bound and report coverage probabilities and average lengths of 95% confidence intervals, for samples sizes  $n \in \{100, 1000, 5000, 10000\}$ , out of 1000 Monte Carlo replications in Table 6. I use empirical means in the first-step estimation of conditional mean treatment responses and propensity scores. I construct the confidence intervals using original and debiased moment conditions with and without cross-fitting. Confidence intervals constructed using original moments are invalid, and as expected, show undercoverage. However, confidence intervals obtained using debiased moment conditions show good coverage even with small sample size. I also report the results when true values of nuisance parameters are used to construct the confidence intervals. In that case, the coverage probability is around 0.95 for both original and debiased moments, as expected.

## 7 Conclusion

In this paper, I consider identification and inference of the welfare gain that results from switching from one policy to another policy. Understanding how much the welfare gain is under different assumptions on the unobservables allows policymakers to make informed decisions about how to choose between alternative treatment assignment policies. I use tools from theory of random sets to obtain the identified set of this parameter. I then employ orthogonalized moment conditions for the estimation and inference of these bounds. I illustrate the usefulness of the analysis by considering hypothetical policies with experimental data from the National JTPA study. I conduct Monte Carlo simulations to assess the finite sample performance of the estimators.

Table 6: 95% confidence interval for worst-case lower bound

Sample size	Original moment		Debiased moment	
	Coverage	Average length	Coverage	Average length
when first-step is estimated with empirical means				
without cross-fitting				
100	0.80	13976	0.94	21316
1000	0.79	4454	0.95	6797
5000	0.78	1995	0.94	3045
10000	0.80	1412	0.96	2154
with cross-fitting ( $L = 2$ )				
100	0.79	14180	0.94	21316
1000	0.78	4462	0.95	6797
5000	0.78	1996	0.94	3045
10000	0.80	1412	0.96	2154
when true values of nuisance parameters are used				
100	0.95	14008	0.94	21316
1000	0.94	4449	0.95	6797
5000	0.95	1991	0.94	3045
10000	0.95	1408	0.96	2154

Note: number of Monte Carlo replications is 1000

## A Random Set Theory

In this appendix, I introduce some definitions and theorems from random set theory that are used throughout the paper. See [Molchanov \(2017\)](#) and [Molchanov and Molinari \(2018\)](#) for more detailed treatment of random set theory. Let  $(\Omega, \mathfrak{A}, P)$  be a complete probability space and  $\mathcal{F}$  be the family of closed subsets of  $\mathbb{R}^d$ .

**Definition 1** (Random closed set). A map  $X : \Omega \rightarrow \mathcal{F}$  is called a random closed set if, for every compact set  $K$  in  $\mathbb{R}^d$ ,

$$\{\omega \in \Omega : X(\omega) \cap K \neq \emptyset\} \in \mathfrak{A}. \quad (83)$$

**Definition 2** (Selection). A random vector  $\xi$  with values in  $\mathbb{R}^d$  is called a (measurable) selection of  $X$  if  $\xi(\omega) \in X(\omega)$  for almost all  $\omega \in \Omega$ . The family of all selections of  $X$  is denoted by  $\mathcal{S}(X)$ .

**Definition 3** (Integrable selection). Let  $L^1 = L^1(\Omega; \mathbb{R}^d)$  denote the space of  $\mathfrak{A}$ -measurable random vectors with values in  $\mathbb{R}^d$  such that the  $L^1$ -norm  $\|\xi\|_1 = E[\|\xi\|]$  is finite. If  $X$  is a random closed set in  $\mathbb{R}^d$ , then the family of all integrable selections of  $X$  is given by

$$\mathcal{S}^1(X) = \mathcal{S}(X) \cap L^1. \quad (84)$$

**Definition 4** (Integrable random sets). A random closed set  $X$  is called integrable if  $\mathcal{S}^1(X) \neq \emptyset$ .

**Definition 5** (Selection (or Aumann) expectation). The selection (or Aumann) expectation of  $X$  is the closure of the set of all expectations of integrable selections, i.e.

$$\mathbb{E}[X] = cl\left\{\int_{\Omega} \xi dP : \xi \in \mathcal{S}^1(X)\right\}. \quad (85)$$

Note that I use  $\mathbb{E}[\cdot]$  for the Aumann expectation and reserve  $E[\cdot]$  for the expectation of random variables and random vectors.

**Definition 6** (Support function). Let  $K \subset \mathbb{R}^d$  be a convex set. The support function of a set  $K$  is given by

$$s(v, K) = \sup_{x \in K} \langle v, x \rangle, \quad v \in \mathbb{R}^d. \quad (86)$$

**Theorem 5** (Theorem 3.11 in [Molchanov and Molinari \(2018\)](#)). *If an integrable random set  $X$  is defined on a nonatomic probability space, or if  $X$  is almost surely convex, then*

$$E[s(v, X)] = s(v, \mathbb{E}[X]), \quad v \in \mathbb{R}^d. \quad (87)$$



## B Proofs and Useful Lemmas

### Proof of Lemma 1

By the definition of selection expectation, we have  $E[(Y_1, Y_0)'|X] \in \mathbb{E}[\mathcal{Y}_1 \times \mathcal{Y}_0|X]$ . Then by the definition of support function and Theorem 5, for any  $v \in \mathbb{R}^2$ , we have

$$\begin{aligned} v'E[(Y_1, Y_0)'|X] &\leq s(v, \mathbb{E}[\mathcal{Y}_1 \times \mathcal{Y}_0|X]) \\ &= E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]. \end{aligned} \quad (88)$$

For any  $v \in \mathbb{R}^2$ , we can write

$$\begin{aligned} -v'E[(Y_1, Y_0)'|X] &\leq s(-v, \mathbb{E}[\mathcal{Y}_1 \times \mathcal{Y}_0|X]) \\ &= E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]. \end{aligned} \quad (89)$$

Thus, we also have

$$v'E[(Y_1, Y_0)'|X] \geq -E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]. \quad (90)$$

■

### Proof of Theorem 1

We write  $\Delta(X) \equiv E[Y_1 - Y_0|X] = v^{*'}E[(Y_1, Y_0)'|X]$  for  $v^* = (1, -1)'$ . By Lemma 1, we have

$$\underline{\Delta}(X) = -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] \leq \Delta(X) \leq E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] = \bar{\Delta}(X) \quad a.s. \quad (91)$$

Since  $\delta(X) - \delta^*(X)$  can take values in  $\{-1, 0, 1\}$ , we consider two cases: (i)  $\delta(X) - \delta^*(X) = 1$  and (ii)  $\delta(X) - \delta^*(X) = -1$ . When (i)  $\delta(X) - \delta^*(X) = 1$ , the upper bound on  $\Delta(X) \cdot (\delta(X) - \delta^*(X))$  is  $\bar{\Delta}(X)$ . When (ii)  $\delta(X) - \delta^*(X) = -1$ , the upper bound on  $\Delta(X) \cdot (\delta(X) - \delta^*(X))$  is  $-\underline{\Delta}(X)$ . Hence, the upper bound on  $E[\Delta(X) \cdot (\delta(X) - \delta^*(X))]$  should be

$$\beta_u = E[\bar{\Delta}(X) \cdot \theta_{10}(X) - \underline{\Delta}(X) \cdot \theta_{01}(X)]. \quad (92)$$

Similarly, the lower bound on  $E[\Delta(X) \cdot (\delta(X) - \delta^*(X))]$  should be

$$\beta_l = E[\underline{\Delta}(X) \cdot \theta_{10}(X) - \bar{\Delta}(X) \cdot \theta_{01}(X)]. \quad (93)$$

■

**Lemma 7.** Suppose  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  is of the following form:

$$\mathcal{Y}_1 \times \mathcal{Y}_0 = \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0, \end{cases} \quad (94)$$

where  $Y$  is a random variable and each of  $Y_{L,0}, Y_{U,0}, Y_{L,1}$ , and  $Y_{U,1}$  can be a constant or a random variable. Let  $v^* = (1, -1)'$ ,  $v_1 = (1, 0)'$ , and  $v_0 = (0, 1)'$ . Then, we have

$$\begin{aligned} E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E[Y|D = 1, X] \cdot P(D = 1|X) + E[Y_{U,1}|D = 0, X] \cdot P(D = 0|X), \\ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E[Y|D = 1, X] \cdot P(D = 1|X) + E[Y_{L,1}|D = 0, X] \cdot P(D = 0|X), \\ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E[Y_{U,0}|D = 1, X] \cdot P(D = 1|X) + E[Y|D = 0, X] \cdot P(D = 0|X), \\ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E[Y_{L,0}|D = 1, X] \cdot P(D = 1|X) + E[Y|D = 0, X] \cdot P(D = 0|X), \\ E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= (E[Y|D = 1, X] - E[Y_{L,0}|D = 1, X]) \cdot P(D = 1|X) \\ &\quad + (E[Y_{U,1}|D = 0, X] - E[Y|D = 0, X]) \cdot P(D = 0|X), \\ -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= (E[Y|D = 1, X] - E[Y_{U,0}|D = 1, X]) \cdot P(D = 1|X) \\ &\quad + (E[Y_{L,1}|D = 0, X] - E[Y|D = 0, X]) \cdot P(D = 0|X). \end{aligned}$$

*Proof.* We have

$$\begin{aligned}
E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_1 | X\right] \\
&= E[Y|D = 1, X] \cdot P(D = 1|X) + E[Y_{U,1}|D = 0, X] \cdot P(D = 0|X),
\end{aligned}$$

$$\begin{aligned}
-E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= -E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} -y_1 | X\right] \\
&= E\left[\inf_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_1 | X\right] \\
&= E[Y|D = 1, X] \cdot P(D = 1|X) + E[Y_{L,1}|D = 0, X] \cdot P(D = 0|X),
\end{aligned}$$

$$\begin{aligned}
E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_0 | X\right] \\
&= E[Y_{U,0}|D = 1, X] \cdot P(D = 1|X) + E[Y|D = 0, X] \cdot P(D = 0|X),
\end{aligned}$$

$$\begin{aligned}
-E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= -E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} -y_0 | X\right] \\
&= E\left[\inf_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_0 | X\right] \\
&= E[Y_{L,0}|D = 1, X] \cdot P(D = 1|X) + E[Y|D = 0, X] \cdot P(D = 0|X),
\end{aligned}$$

$$\begin{aligned}
E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_1 - y_0 | X\right] \\
&= (E[Y|D = 1, X] - E[Y_{L,0}|D = 1, X]) \cdot P(D = 1|X) \\
&\quad + (E[Y_{U,1}|D = 0, X] - E[Y|D = 0, X]) \cdot P(D = 0|X), \\
\\
-E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] &= -E\left[\sup_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} -y_1 + y_0 | X\right] \\
&= E\left[\inf_{(y_1, y_0) \in \begin{cases} \{Y\} \times [Y_{L,0}, Y_{U,0}] & \text{if } D = 1, \\ [Y_{L,1}, Y_{U,1}] \times \{Y\} & \text{if } D = 0. \end{cases}} y_1 - y_0 | X\right] \\
&= (E[Y|D = 1, X] - E[Y_{U,0}|D = 1, X]) \cdot P(D = 1|X) \\
&\quad + (E[Y_{L,1}|D = 0, X] - E[Y|D = 0, X]) \cdot P(D = 0|X). \quad \blacksquare
\end{aligned}$$

### Proof of Corollary 1

By setting  $Y_{L,1} = Y_{L,0} = \underline{y}$  and  $Y_{U,1} = Y_{U,0} = \bar{y}$  in Lemma 7, we have

$$E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] = (\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X)), \quad (95)$$

$$-E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] = (\eta(1, X) - \bar{y}) \cdot p(X) + (\underline{y} - \eta(0, X)) \cdot (1 - p(X)). \quad (96)$$

Plugging these in, the result follows from Theorem 1.  $\blacksquare$

### Proof of Corollary 2

By setting  $Y_{L,0} = \underline{y}$ ,  $Y_{U,0} = Y_{L,1} = Y$ , and  $Y_{U,1} = \bar{y}$  in Lemma 7, we have

$$E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] = (\eta(1, X) - \underline{y}) \cdot p(X) + (\bar{y} - \eta(0, X)) \cdot (1 - p(X)), \quad (97)$$

$$-E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X] = 0. \quad (98)$$

Plugging these in, the result follows from Theorem 1.  $\blacksquare$

## Proof of Lemma 2

By the definition of selection expectation, we have  $E[(Y_1, Y_0)'|X, Z] \in \mathbb{E}[\mathcal{Y}_1 \times \mathcal{Y}_0|X, Z]$ . By arguments that appear in Lemma 1, for any  $v \in \mathbb{R}^2$  and for all  $z \in \mathcal{Z}$ , we have

$$-E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \leq v'E[(Y_1, Y_0)'|X, Z = z] \leq E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z]. \quad (99)$$

Assumption 2 implies that

$$E[Y_d|X, Z] = E[Y_d|X], \quad d = 0, 1. \quad (100)$$

Hence, for all  $z \in \mathcal{Z}$ , the following holds:

$$-E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \leq v'E[(Y_1, Y_0)'|X] \leq E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z]. \quad (101)$$

We therefore have

$$\sup_{z \in \mathcal{Z}} \{ -E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} \leq v'E[(Y_1, Y_0)'|X] \leq \inf_{z \in \mathcal{Z}} \{ E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \}. \quad (102)$$

## Proof of Theorem 2

By Lemma 2, we have

$$\sup_{z \in \mathcal{Z}} \{ -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} \leq \Delta(X) \leq \inf_{z \in \mathcal{Z}} \{ E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] \} \quad a.s. \quad (103)$$

The remaining part of the proof is the same as that of Theorem 1. ■

### Proof of Corollary 3

The statements in Lemma 7 still hold when we condition on an additional variable  $Z$ . Hence, by setting  $Y_{L,1} = Y_{L,0} = \underline{y}$  and  $Y_{U,1} = Y_{U,0} = \bar{y}$  in Lemma 7, we have

$$E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z] = (\eta(1, X, z) - \underline{y}) \cdot p(X, z) + (\bar{y} - \eta(0, X, z)) \cdot (1 - p(X, z)), \quad (104)$$

$$-E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z] = (\eta(1, X, z) - \bar{y}) \cdot p(X, z) + (\underline{y} - \eta(0, X, z)) \cdot (1 - p(X, z)), \quad (105)$$

for all  $z \in \mathcal{Z}$ . Plugging these in, the result follows from Theorem 2. ■

### Proof of Corollary 4

The statements in Lemma 7 still hold when we condition on an additional variable  $Z$ . Hence, by setting  $Y_{L,0} = \underline{y}$ ,  $Y_{U,0} = Y_{L,1} = Y$ , and  $Y_{U,1} = \bar{y}$  in Lemma 7, we have

$$E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z] = (\eta(1, X, z) - \underline{y}) \cdot p(X, z) + (\bar{y} - \eta(0, X, z)) \cdot (1 - p(X, z)), \quad (106)$$

$$-E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z] = 0, \quad (107)$$

for all  $z \in \mathcal{Z}$ . Plugging these in, the result follows from Theorem 2. ■

### Proof of Lemma 4

By the definition of selection expectation, we have  $E[(Y_1, Y_0)' | X, Z] \in \mathbb{E}[\mathcal{Y}_1 \times \mathcal{Y}_0 | X, Z]$ . By arguments that appear in Lemma 1, for any  $v \in \mathbb{R}_+^2$  and for all  $z \in \mathcal{Z}$ , we have

$$-E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z] \leq v' E[(Y_1, Y_0)' | X, Z = z] \leq E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z]. \quad (108)$$

By Assumption 4, the following holds for all  $z \in \mathcal{Z}$ :

$$\sup_{z_1 \leq z} \{-E[s(-v, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_1]\} \leq v' E[(Y_1, Y_0)' | X, Z = z] \leq \inf_{z_2 \geq z} \{E[s(v, \mathcal{Y}_1 \times \mathcal{Y}_0) | X, Z = z_2]\}. \quad (109)$$

By replacing  $v$  with  $v_1 = (1, 0)'$  and  $v_0 = (0, 1)'$  and integrating everything with respect to  $Z$ , we obtain the following:

$$E[Y_1|X] \geq \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \left\{ -E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z_1] \right\} \right), \quad (110)$$

$$E[Y_1|X] \leq \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \left\{ E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z_2] \right\} \right), \quad (111)$$

$$E[Y_0|X] \geq \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \sup_{z_1 \leq z} \left\{ -E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z_1] \right\} \right), \quad (112)$$

$$E[Y_0|X] \leq \sum_{z \in \mathcal{Z}} P(Z = z) \cdot \left( \inf_{z_2 \geq z} \left\{ E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z_2] \right\} \right). \quad (113)$$

Then, the upper bound in (38) can be obtained by subtracting the lower bound on  $E[Y_0|X]$  (112) from the upper bound on  $E[Y_1|X]$  (111). Similarly, the lower bound in (38) can be obtained by subtracting the upper bound on  $E[Y_0|X]$  (113) from the lower bound on  $E[Y_1|X]$  (110). ■

### Proof of Theorem 3

Bounds on  $\Delta(X)$  is derived in Lemma 4. The remaining part of the proof is the same as that of Theorem 1. ■

### Proof of Corollary 5

The statements in Lemma 7 still hold when we condition on an additional variable  $Z$ . Hence, by setting  $Y_{L,1} = Y_{L,0} = \underline{y}$  and  $Y_{U,1} = Y_{U,0} = \bar{y}$  in Lemma 7, for all  $z \in \mathcal{Z}$ , we have

$$E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)), \quad (114)$$

$$-E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \eta(1, X, z) \cdot p(X, z) + \underline{y} \cdot (1 - p(X, z)), \quad (115)$$

$$E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \bar{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)), \quad (116)$$

$$-E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)). \quad (117)$$

Plugging these in, the result follows from Theorem 3. ■

## Proof of Corollary 6

The statements in Lemma 7 still hold when we condition on an additional variable  $Z$ . Hence, by setting  $Y_{L,0} = \underline{y}$ ,  $Y_{U,0} = Y_{L,1} = Y$ , and  $Y_{U,1} = \bar{y}$  in Lemma 7, for all  $z \in \mathcal{Z}$ , we also have

$$E[s(v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \eta(1, X, z) \cdot p(X, z) + \bar{y} \cdot (1 - p(X, z)), \quad (118)$$

$$-E[s(-v_1, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = E[Y|X, Z = z], \quad (119)$$

$$E[s(v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = E[Y|X, Z = z], \quad (120)$$

$$-E[s(-v_0, \mathcal{Y}_1 \times \mathcal{Y}_0)|X, Z = z] = \underline{y} \cdot p(X, z) + \eta(0, X, z) \cdot (1 - p(X, z)), \quad (121)$$

Plugging these in, the result follows from Theorem 3. ■

## Proof of Lemma 5

For  $0 \leq \tau \leq 1$ , let

$$F_\tau = (1 - \tau)F_0 + \tau G_w^j, \quad (122)$$

where  $F_0$  is the true distribution of  $F$  and  $G_w^j$  is a family of distributions approaching the CDF of a constant  $w$  as  $j \rightarrow \infty$ . Let  $F_0$  be absolutely continuous with pdf  $f_0(w) = f_0(y, d, x)$ . Let the marginal, conditional, and joint distributions and densities under  $F_0$  be denoted by  $F_0(x), F_0(d|x), F_0(y|d, x), F_0(d, x)$  and  $f_0(x), f_0(d|x), f_0(y|d, x), f_0(d, x)$ , etc. and the expectations under  $F_0$  be denoted by  $E_0$ . As in Ichimura and Newey (2017), let

$$G_w^j(\tilde{w}) = E[1\{w_i \leq \tilde{w}\}\varphi(w_i)], \quad (123)$$

where  $\varphi(w_i)$  is a bounded function with  $E[\varphi(w_i)] = 1$ . This  $G_w^j(\tilde{w})$  will approach the cdf of the constant  $\tilde{w}$  as  $\varphi(w)f_0(w)$  approaches a spike at  $\tilde{w}$ . For small enough  $\tau$ ,  $F_\tau$  will be a cdf with pdf  $f_\tau$  that is given by

$$f_\tau(\tilde{w}) = f_0(\tilde{w})[1 - \tau + \tau\varphi(w)] = f_0(\tilde{w})(1 + \tau S(w)), S(w) = \varphi(w) - 1. \quad (124)$$

Let the marginal, conditional, and joint distributions and densities under  $F_\tau$  be similarly denoted by  $F_\tau(x), F_\tau(d|x), F_\tau(y|d, x), F_\tau(d, x)$  and  $f_\tau(x), f_\tau(d|x), f_\tau(y|d, x), f_\tau(d, x)$ , etc. and the expectations under  $F_\tau$  be denoted by  $E_\tau$ . By Ichimura and Newey (2017)'s Lemma A1, we have



$$\frac{d}{d\tau}E_\tau[Y|D=d, X=x] = E_0[\{Y - E_0[Y|D=d, X=x]\}\varphi(W)|D=d, X=x] \quad (125)$$

and

$$\frac{d}{d\tau}E_\tau[1\{D=d\}|X=x] = E_0[\{1\{D=d\} - E_0[1\{D=d\}|X=x]\}\varphi(W)|X=x]. \quad (126)$$

The influence function can be calculated as

$$\phi(w, \beta, \gamma) = \lim_{j \rightarrow \infty} \left[ \frac{d}{d\tau}E_\tau[m(w_i, \beta, \gamma(F_\tau))] \right]_{\tau=0}. \quad (127)$$

We first denote the conditional mean treatment response and the propensity score under  $F_\tau$  by

$$\eta_\tau(d, x) \equiv \int y dF_\tau(y|d, x), \quad (128)$$

and

$$p_\tau(x) \equiv \int 1\{d=1\} dF_\tau(d|x). \quad (129)$$

Then, by the chain rule, we have

$$\begin{aligned} \frac{d}{d\tau}E_\tau[m(w_i, \beta, \gamma(F_\tau))] &= \frac{d}{d\tau}E_\tau[m(w_i, \beta, \gamma(F_0))] + \frac{d}{d\tau}E_0[m(w_i, \beta, \gamma(F_\tau))] \\ &= \frac{d}{d\tau} \left[ \int \left( ((\eta_0(1, x) - \bar{y})p_0(x) + (\underline{y} - \eta_0(0, x))(1 - p_0(x)))\theta_{10}(x) \right. \right. \\ &\quad \left. \left. - ((\eta_0(1, x) - \underline{y})p_0(x) + (\bar{y} - \eta_0(0, x))(1 - p_0(x)))\theta_{01}(x)) - \beta \right) dF_\tau(x) \right] \\ &\quad + \frac{d}{d\tau} \left[ \int \left( ((\eta_0(1, x) - \bar{y})p_\tau(x) + (\underline{y} - \eta_0(0, x))(1 - p_\tau(x)))\theta_{10}(x) \right. \right. \\ &\quad \left. \left. - ((\eta_0(1, x) - \underline{y})p_\tau(x) + (\bar{y} - \eta_0(0, x))(1 - p_\tau(x)))\theta_{01}(x)) - \beta \right) dF_0(x) \right] \\ &\quad + \frac{d}{d\tau} \left[ \int \left( ((\eta_\tau(1, x) - \bar{y})p_0(x) + (\underline{y} - \eta_\tau(0, x))(1 - p_0(x)))\theta_{10}(x) \right. \right. \\ &\quad \left. \left. - ((\eta_\tau(1, x) - \underline{y})p_0(x) + (\bar{y} - \eta_\tau(0, x))(1 - p_0(x)))\theta_{01}(x)) - \beta \right) dF_0(x) \right]. \end{aligned}$$

First, we have

$$\begin{aligned} \frac{d}{d\tau} E_\tau[m(w_i, \beta, \gamma(F_0))] &= \int \left( (\eta_0(1, x) - \bar{y})p_0(x) + (\underline{y} - \eta_0(0, x))(1 - p_0(x)) \right) \theta_{10}(x) \\ &\quad - \left( (\eta_0(1, x) - \underline{y})p_0(x) + (\bar{y} - \eta_0(0, x))(1 - p_0(x)) \right) \theta_{01}(x) \right) dG(x) - \beta. \end{aligned}$$

Next, we want to find  $\frac{d}{d\tau} E_0[m(w_i, \beta, \gamma(F_\tau))]$ . In order to do that, first note that we have

$$\begin{aligned} &\frac{d}{d\tau} \int \theta(x) \eta_\tau(d, x) f_0(d|x) f_0(x) dx \\ &= \int \theta(x) \frac{d}{d\tau} [\eta_\tau(d, x)] f_0(d|x) f_0(x) dx \\ &= \int \theta(x) E_0[\{Y - \eta_0(d, x)\} \varphi(W) | D = d, X = x] f_0(d|x) f_0(x) dx \\ &= \int \theta(x) \left[ \int \{y - \eta_0(d, x)\} \frac{g(y, d, x)}{f_0(y, d, x)} f_0(y|d, x) dy \right] f_0(d|x) f_0(x) dx \\ &= \int \theta(x) \left[ \int \{y - \eta_0(d, x)\} \frac{g(y, d, x)}{f_0(y, d, x)} \frac{f_0(y, d, x)}{f_0(d, x)} dy \right] \frac{f_0(d, x)}{f_0(x)} f_0(x) dx \\ &= \int \theta(x) \left[ \int \{y - \eta_0(d, x)\} g(y, d, x) dy \right] dx \\ &= \int \theta(x) \{y - \eta_0(d, x)\} g(y, d, x) dy dx. \end{aligned}$$

The second equality follows from equation (125). The third equality follows from choosing  $\varphi(w)$  to be a ratio of a sharply peaked pdf to the true density:

$$\varphi(\tilde{w}) = \frac{g(\tilde{w}) 1(f_0(\tilde{w}) \geq 1/j)}{f_0(\tilde{w})}, \quad (130)$$

where as in Ichimura and Newey (2017),  $g(w)$  is specified as follows. Letting  $K(u)$  be a pdf that is symmetric around zero, has bounded support, and is continuously differentiable of all orders with bounded derivatives, we let

$$g(\tilde{w}) = \prod_{l=1}^r \kappa_l^j(\tilde{w}_l), \kappa_l^j(\tilde{w}_l) = \frac{j K((w_l - \tilde{w}_l)j)}{j \int K((w_l - \tilde{w}_l)j) d\mu_l(\tilde{w}_l)}. \quad (131)$$

Hence, we obtain

$$\begin{aligned}
& \frac{d}{d\tau} \left[ \int \left( ((\eta_\tau(1, x) - \bar{y})p_0(x) + (\underline{y} - \eta_\tau(0, x))(1 - p_0(x)))\theta_{10}(x) \right. \right. \\
& \quad \left. \left. - ((\eta_\tau(1, x) - \underline{y})p_0(x) + (\bar{y} - \eta_\tau(0, x))(1 - p_0(x)))\theta_{01}(x)) - \beta \right) dF_0(x) \right] \\
&= \int (\theta_{10}(x) - \theta_{01}(x)) \{y - \eta_0(1, x)\} g(y, 1, x) dy dx \\
& \quad - \int (\theta_{10}(x) - \theta_{01}(x)) \{y - \eta_0(0, x)\} g(y, 0, x) dy dx.
\end{aligned}$$

With the similar argument, but using equation (126), we also have

$$\begin{aligned}
& \frac{d}{d\tau} \int \theta(x) p_\tau(x) f_0(x) dx \\
&= \int \theta(x) \frac{d}{d\tau} [p_\tau(x)] f_0(x) dx \\
&= \int \theta(x) E_0[\{1\{D = 1\} - p_0(x)\} \varphi(W) | X = x] f_0(x) dx \\
&= \int \theta(x) \left[ \int \{1\{d = 1\} - p_0(x)\} \frac{g(y, d, x)}{f_0(y, d, x)} f_0(y, d|x) dy dd \right] f_0(x) dx \\
&= \int \theta(x) \left[ \int \{1\{d = 1\} - p_0(x)\} \frac{g(y, d, x)}{f_0(y, d, x)} \frac{f_0(y, d, x)}{f_0(x)} dy dd \right] f_0(x) dx \\
&= \int \theta(x) \{1\{d = 1\} - p_0(x)\} g(y, d, x) dy dddx.
\end{aligned}$$

Hence,

$$\begin{aligned}
& \frac{d}{d\tau} \left[ \int \left( ((\eta_0(1, x) - \bar{y})p_\tau(x) + (\underline{y} - \eta_0(0, x))(1 - p_\tau(x)))\theta_{10}(x) \right. \right. \\
& \quad \left. \left. - ((\eta_0(1, x) - \underline{y})p_\tau(x) + (\bar{y} - \eta_0(0, x))(1 - p_\tau(x)))\theta_{01}(x)) - \beta \right) dF_0(x) \right] \\
&= \int (\theta_{10}(x) - \theta_{01}(x)) (\eta_0(1, x) + \eta_0(0, x) - (\underline{y} + \bar{y})) \{1\{d = 1\} - p_0(x)\} g(y, d, x) dy dddx
\end{aligned}$$

Therefore, as  $j \rightarrow \infty$ , since  $\eta(1, x)$ ,  $\eta(0, x)$ , and  $p(x)$  are continuous at  $x$ , we obtain

$$\begin{aligned}\phi(w, \beta, \gamma) = & \left( (\eta_0(1, x) - \bar{y})p_0(x) + (\underline{y} - \eta_0(0, x))(1 - p_0(x)) \right) \theta_{10}(x) \\ & - \left( (\eta_0(1, x) - \underline{y})p_0(x) + (\bar{y} - \eta_0(0, x))(1 - p_0(x)) \right) \theta_{01}(x) - \beta \\ & + (\theta_{10}(x) - \theta_{01}(x))(\eta_0(1, x) + \eta_0(0, x) - (\underline{y} + \bar{y}))\{1\{d = 1\} - p_0(x)\} \\ & + (\theta_{10}(x) - \theta_{01}(x))\{y - \eta_0(1, x)\}^d \{-(y - \eta_0(0, x))\}^{1-d}. \quad \blacksquare\end{aligned}$$

## Proof of Theorem 4

Let

$$\hat{\psi}(\beta_l) = \frac{1}{n} \sum_{k=1}^L \sum_{i \in \mathcal{I}_k} \psi(w_i, \beta_l, \hat{\gamma}_k). \quad (132)$$

First we show that

$$\sqrt{n} \hat{\psi}(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(w_i, \beta_0, \gamma_0) + o_p(1) \quad (133)$$

holds. Under Assumption 6 (i), (ii), and (iii), the result follows. Following CEINR, we provide a sketch of the argument. Let

$$\hat{\Delta}_{ik} \equiv \psi(w_i, \beta_0, \hat{\gamma}_k) - \bar{\psi}(\hat{\gamma}_k) - \psi(w_i, \beta_0, \gamma_0), \quad (134)$$

and

$$\bar{\Delta}_k \equiv \frac{1}{n} \sum_{i \in \mathcal{I}_k} \hat{\Delta}_{ik}. \quad (135)$$

Let  $n_k$  be the number of observations with  $i \in \mathcal{I}_k$  and  $W_k$  denote a vector of all observations  $w_i$  for  $i \notin \mathcal{I}_k$ . Note that for any  $i, j \in \mathcal{I}_k$ ,  $i \neq j$ , we have  $E[\hat{\Delta}_{ik} \hat{\Delta}_{jk} | W_k] = E[\hat{\Delta}_{ik} | W_k] E[\hat{\Delta}_{jk} | W_k] = 0$  since by construction  $E[\hat{\Delta}_{ik} | W_k] = 0$ . By Assumption 6 (i),

$$E[\bar{\Delta}_k^2 | W_k] = \frac{1}{n^2} \sum_{i \in \mathcal{I}_k} E[\hat{\Delta}_{ik}^2 | W_k] \leq \frac{n_k}{n^2} \int \{\psi(w, \beta_0, \hat{\gamma}_k) - \psi(w, \beta_0, \gamma_0)\}^2 F_0(dw) = o_p(n_k/n^2). \quad (136)$$

This implies that, for each  $k$ , we have  $\bar{\Delta}_k = o_p(\sqrt{n_k}/n)$ . Then it follows that

$$\sqrt{n} \left[ \hat{\psi}(\beta_0) - \frac{1}{n} \sum_{i=1}^n \psi(w_i, \beta_0, \gamma_0) - \frac{1}{n} \sum_{k=1}^L n_k \bar{\psi}(\hat{\gamma}_k) \right] = \sqrt{n} \sum_{k=1}^L \bar{\Delta}_k = o_p(\sqrt{n_k}/n) \xrightarrow{p} 0. \quad (137)$$

By Assumption 6 (ii) and (iii), we have

$$\sqrt{n}|\bar{\psi}(\hat{\gamma}_k)| \leq \sqrt{n}C\|\hat{\gamma}_k - \gamma_0\|^2 \xrightarrow{p} 0. \quad (138)$$

Then (133) follows by the triangle inequality. Since (133) holds and  $\{w_i\}_{i=1}^n$  are i.i.d., by central limit theorem

$$\sqrt{n}\hat{\psi}(\beta_0) \xrightarrow{d} N(0, \Omega), \quad (139)$$

where  $\Omega = E[\psi(w_i, \beta_0, \gamma_0)^2]$ . The rest of the proof is standard as in Newey and McFadden (1994) and we provide a sketch of the argument. Let  $M = E[\frac{\partial \psi(w, \beta, \gamma_0)}{\partial \beta}|_{\beta=\beta_0}]$  and  $\hat{M} = \frac{\partial \hat{\psi}(\hat{\beta})}{\partial \beta}$ . The first order condition is

$$0 = \hat{M}\hat{\psi}(\hat{\beta}). \quad (140)$$

We expand  $\hat{\psi}(\hat{\beta})$  around  $\beta_0$  to obtain

$$\hat{\psi}(\hat{\beta}) = \hat{\psi}(\beta_0) + \bar{M}(\hat{\beta} - \beta_0), \quad (141)$$

where  $\bar{M} = \frac{\partial \hat{\psi}(\beta_u)}{\partial \beta}$  and  $\bar{\beta}$  is the mean value. Substituting this back into the first order condition, we get

$$0 = \hat{M}\hat{\psi}(\beta_0) + \hat{M}\bar{M}(\hat{\beta} - \beta_0). \quad (142)$$

Solving this for  $\hat{\beta} - \beta_0$  and multiplying by  $\sqrt{n}$ , we obtain

$$\sqrt{n}(\hat{\beta} - \beta_0) = -(\hat{M}\bar{M})^{-1}\hat{M}\sqrt{n}\hat{\psi}(\beta_0). \quad (143)$$

We also have  $\hat{M} \xrightarrow{p} M$  and  $\bar{M} \xrightarrow{p} M$  and by the continuous mapping theorem,

$$-(\hat{M}\bar{M})^{-1}\hat{M} \xrightarrow{p} -M^{-1}. \quad (144)$$

Then, by the Slutsky theorem,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} -M^{-1}N(0, \Omega) = N(0, M^{-2}\Omega). \quad (145)$$

In our case,  $M = E[\frac{\partial \psi(w, \beta, \gamma_0)}{\partial \beta}|_{\beta=\beta_0}] = 1$  and so the asymptotic variance is  $\Omega$ . Finally, CEINR showed that Assumption 6 (iii) insures that  $\hat{V} \xrightarrow{p} V$ . ■

## Proof of Lemma 6

The proof is similar to that of Lemma 5. Since we have an additional variable  $Z \in \{0, 1\}$ , we make slight adjustments. By the chain rule, we have

$$\begin{aligned}
\frac{d}{d\tau} E_\tau[m(w_i, \beta, \gamma(F_\tau))] &= \frac{d}{d\tau} E_\tau[m(w_i, \beta, \gamma(F_0))] + \frac{d}{d\tau} E_0[m(w_i, \beta, \gamma(F_\tau))] \\
&= \frac{d}{d\tau} \left[ \int \left( (\eta_0(1, x, 1) \cdot p_0(x, 1) + \underline{y} \cdot (1 - p_0(x, 1)) - \bar{y} \cdot p_0(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{10}(x) \right. \right. \\
&\quad \left. \left. - (\eta_0(1, x, 1) \cdot p_0(x, 1) + \bar{y} \cdot (1 - p_0(x, 1)) - \underline{y} \cdot p_0(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{01}(x) - \beta \right) dF_\tau(x) \right] \\
&\quad + \frac{d}{d\tau} \left[ \int \left( (\eta_0(1, x, 1) \cdot p_\tau(x, 1) + \underline{y} \cdot (1 - p_\tau(x, 1)) - \bar{y} \cdot p_\tau(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_\tau(x, 0))) \cdot \theta_{10}(x) \right. \right. \\
&\quad \left. \left. - (\eta_0(1, x, 1) \cdot p_\tau(x, 1) + \bar{y} \cdot (1 - p_\tau(x, 1)) - \underline{y} \cdot p_\tau(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_\tau(x, 0))) \cdot \theta_{01}(x) - \beta \right) dF_0(x) \right] \\
&\quad + \frac{d}{d\tau} \left[ \int \left( (\eta_\tau(1, x, 1) \cdot p_0(x, 1) + \underline{y} \cdot (1 - p_0(x, 1)) - \bar{y} \cdot p_0(x, 0) - \eta_\tau(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{10}(x) \right. \right. \\
&\quad \left. \left. - (\eta_\tau(1, x, 1) \cdot p_0(x, 1) + \bar{y} \cdot (1 - p_0(x, 1)) - \underline{y} \cdot p_0(x, 0) - \eta_\tau(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{01}(x) - \beta \right) dF_0(x) \right].
\end{aligned}$$

By the arguments in the proof of Lemma 5, we have

$$\frac{d}{d\tau} \int \theta(x) \eta_\tau(d, x, z) f_0(x) dx = \int \theta(x) \{y - \eta_0(d, x, z)\} g(y, d, x) dy dx. \quad (146)$$

and

$$\frac{d}{d\tau} \int \theta(x) p_\tau(x, z) f_0(x) dx = \int \theta(x) \{1\{d = 1\} - p_0(x, z)\} g(y, d, x) dy d x. \quad (147)$$

Hence, we obtain

$$\begin{aligned}
& \frac{d}{d\tau} \left[ \int \left( (\eta_0(1, x, 1) \cdot p_\tau(x, 1) + \underline{y} \cdot (1 - p_\tau(x, 1)) - \bar{y} \cdot p_\tau(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_\tau(x, 0))) \cdot \theta_{10}(x) \right. \right. \\
& \quad \left. \left. - (\eta_0(1, x, 1) \cdot p_\tau(x, 1) + \bar{y} \cdot (1 - p_\tau(x, 1)) - \underline{y} \cdot p_\tau(x, 0) - \eta_0(0, x, 0) \cdot (1 - p_\tau(x, 0))) \cdot \theta_{01}(x) - \beta \right) dF_0(x) \right] \\
&= \int ((\eta_0(1, x, 1) - \underline{y}) \cdot \theta_{10}(x) - (\eta_0(1, x, 1) - \bar{y}) \cdot \theta_{01}(x)) \cdot (d - p_0(x, 1)) \\
& \quad + ((\eta_0(0, x, 0) - \bar{y}) \cdot \theta_{10}(x) - (\eta_0(0, x, 0) - \underline{y}) \cdot \theta_{01}(x)) \cdot (d - p_0(x, 0)) g(y, d, z, x) dy d d d z dx
\end{aligned}$$

Also,

$$\begin{aligned}
& \frac{d}{d\tau} \left[ \int \left( (\eta_\tau(1, x, 1) \cdot p_0(x, 1) + \underline{y} \cdot (1 - p_0(x, 1)) - \bar{y} \cdot p_0(x, 0) - \eta_\tau(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{10}(x) \right. \right. \\
& \quad \left. \left. - (\eta_\tau(1, x, 1) \cdot p_0(x, 1) + \bar{y} \cdot (1 - p_0(x, 1)) - \underline{y} \cdot p_0(x, 0) - \eta_\tau(0, x, 0) \cdot (1 - p_0(x, 0))) \cdot \theta_{01}(x) - \beta \right) dF_0(x) \right] \\
&= \int (\theta_{10}(x) - \theta_{01}(x)) \cdot p_0(x, 1) \cdot (y - \eta_0(1, x, 1)) g(y, d, z, x) dy d d d z dx \\
& \quad - \int (\theta_{10}(x) - \theta_{01}(x)) \cdot (1 - p_0(x, 0)) \cdot (y - \eta_0(0, x, 0)) g(y, d, z, x) dy d d d z dx
\end{aligned}$$

■

## C More General Case

I show how my result can be extended to a more general setting. Let  $\delta : \mathcal{X} \rightarrow [0, 1]$  so that the treatment rules can be randomized treatment rules. Also, let  $w : \mathcal{X} \rightarrow \mathbb{R}_+$  be some weighting function so that the policymaker cares about the weighted average welfare  $E[w(X) \cdot (E[Y_1|X] \cdot \delta(X) + E[Y_0|X] \cdot (1 - \delta(X)))]$  rather than the mean welfare. Then, by letting  $\psi(X) \equiv w(X) \cdot (\delta(X) - \delta^*(X))$ , my object of interest becomes

$$E[\Delta(X) \cdot w(X) \cdot (\delta(X) - \delta^*(X))] = E[\Delta(X) \cdot \psi(X)]. \quad (148)$$

I derive the identification of this parameter in the following theorem.

**Theorem 6** (More general case). *Suppose  $(\mathcal{Y}_1 \times \mathcal{Y}_0) : \Omega \rightarrow \mathcal{F}$  is an integrable random set. Let  $\delta : \mathcal{X} \rightarrow [0, 1]$  and  $\delta^* : \mathcal{X} \rightarrow [0, 1]$  be treatment rules and  $w : \mathcal{X} \rightarrow \mathbb{R}_+$  be a weighting function.*

Also, let  $\psi(X) \equiv w(X) \cdot (\delta(X) - \delta^*(X))$  and  $p = (1, -1)'$ . Then,  $B_I(\delta, \delta^*)$  in (7) is an interval  $[\beta_l, \beta_u]$  where

$$\beta_l = E[\underline{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) \geq 0\} - \bar{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) < 0\}], \quad (149)$$

and

$$\beta_u = E[\bar{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) \geq 0\} - \underline{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) < 0\}], \quad (150)$$

where  $\bar{\Delta}(X) \equiv E[s(v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$  and  $\underline{\Delta}(X) \equiv -E[s(-v^*, \mathcal{Y}_1 \times \mathcal{Y}_0)|X]$ .

*Proof.* The proof is similar to that of Theorem 1. I still have (91) to bound  $\Delta(X)$ . Since  $\psi(X) \in \mathbb{R}$ , I consider two cases: (i)  $\psi(X) \geq 0$  and (ii)  $\psi(X) < 0$ . When (i)  $\psi(X) \geq 0$ , the upper bound on  $\Delta(X) \cdot \psi(X)$  is  $\bar{\Delta}(X) \cdot |\psi(X)|$ . When (ii)  $\psi(X) < 0$ , the upper bound on  $\Delta(X) \cdot \psi(X)$  is  $-\underline{\Delta}(X) \cdot |\psi(X)|$ . Hence, the lower bound should be

$$\beta_l = E[\underline{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) \geq 0\} - \bar{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) < 0\}]. \quad (151)$$

Similarly, the upper bound on  $E[\Delta(X) \cdot \psi(X)]$  should be

$$\beta_u = E[\bar{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) \geq 0\} - \underline{\Delta}(X) \cdot |\psi(X)| \cdot \mathbb{1}\{\psi(X) < 0\}]. \quad (152)$$

■



## D More Details on the Simulation Study

The population welfare gain that results from switching from  $\delta^*(x) = 1\{x \leq 11\}$  to  $\delta(x) = 1\{x \leq 12\}$  is

$$\begin{aligned}\beta &= E[E[Y_1 - Y_0|X] \cdot (\delta(X) - \delta^*(X))] \\ &= P(X = 12) \cdot E[Y_1 - Y_0|X = 12] \\ &= 0.43 \cdot \int_0^1 \{m_1(12, u) - m_0(12, u)\} du \\ &= 1236.\end{aligned}\tag{153}$$

The integration is done using `integrate()` function on R. Given the structure in Section 6, we have

$$E[Y|D = 0, X, Z] = E[Y_0|U > p(X, Z), X, Z] = \frac{1}{1 - p(X, Z)} \int_{p(X, Z)}^1 m_0(X, u) du, \tag{154}$$

$$E[Y|D = 1, X, Z] = E[Y_1|U \leq p(X, Z), X, Z] = \frac{1}{p(X, Z)} \int_0^{p(X, Z)} m_1(X, u) du, \tag{155}$$

$$P(D = 1|X) = 2/3 \cdot p(X, 1) + 1/3 \cdot p(X, 0), \tag{156}$$

$$P(Z = 1|D = 0, X) = \frac{P(D = 0|Z = 1, X)P(Z = 1|X)}{P(D = 0|X)} = \frac{2/3 \cdot (1 - p(X, 1))}{1 - 2/3 \cdot p(X, 1) - 1/3 \cdot p(X, 0)}, \tag{157}$$

$$P(Z = 0|D = 0, X) = \frac{1/3 \cdot (1 - p(X, 0))}{1 - 2/3 \cdot p(X, 1) - 1/3 \cdot p(X, 0)}, \tag{158}$$

$$P(Z = 1|D = 1, X) = \frac{P(D = 1|Z = 1, X)P(Z = 1|X)}{P(D = 1|X)} = \frac{2/3 \cdot p(X, 1)}{2/3 \cdot p(X, 1) + 1/3 \cdot p(X, 0)}, \tag{159}$$

$$P(Z = 0|D = 1, X) = \frac{1/3 \cdot p(X, 0)}{2/3 \cdot p(X, 1) + 1/3 \cdot p(X, 0)}, \tag{160}$$

$$\begin{aligned}E[Y|D = 0, X] &= E[Y|D = 0, X, Z = 1] \cdot P(Z = 1|D = 0, X) \\ &\quad + E[Y|D = 0, X, Z = 0] \cdot P(Z = 0|D = 0, X),\end{aligned}\tag{161}$$

$$\begin{aligned}E[Y|D = 1, X] &= E[Y|D = 1, X, Z = 1] \cdot P(Z = 1|D = 1, X) \\ &\quad + E[Y|D = 1, X, Z = 0] \cdot P(Z = 0|D = 1, X).\end{aligned}\tag{162}$$

Given these quantities, worst-case and IV-worst-case bounds can be calculated similarly.

## E Additional Tables

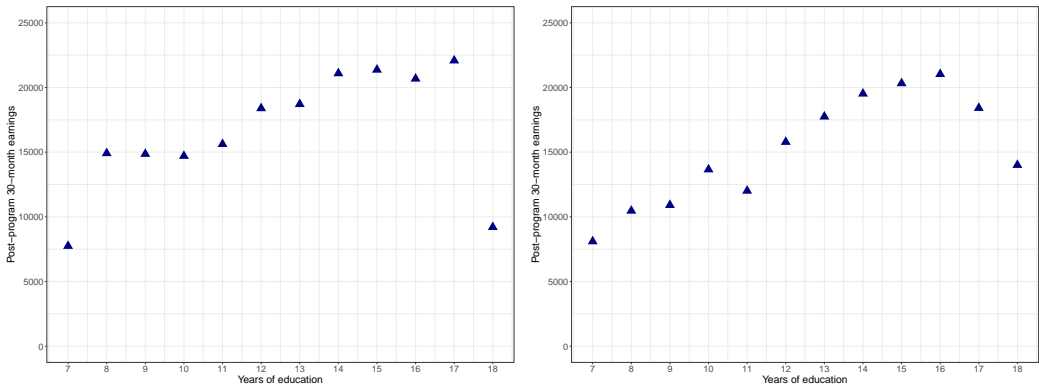
Table 7: Empirical means and distributions when  $X$  is years of education

$X$	7	8	9	10	11	12	13	14	15	16	17	18	Total
sample size	34	616	642	984	1167	3940	660	602	197	260	111	10	9223
$P(X)$	0.004	0.067	0.07	0.107	0.127	0.427	0.072	0.065	0.021	0.028	0.012	0.001	1
$E[Y X]$	7998	12252	12509	14095	13492	16982	18210	20204	20837	20875	20032	11606	
$E[Y D = 1, X]$	7747	14916	14860	14706	15622	18391	18713	21093	21369	20678	22082	9207	
$E[Y D = 0, X]$	8102	10469	10908	13662	12014	15786	17745	19520	20322	21033	18411	14005	
$P(D = 1 X)$	0.294	0.401	0.405	0.415	0.41	0.459	0.48	0.435	0.492	0.446	0.441	0.5	
$E[Y D = 1, Z = 1, X]$	7747	14932	14860	14687	15639	18448	18833	21272	21369	20678	22082	9207	
$E[Y D = 1, Z = 0, X]$	0	11011	14886	17263	14000	13530	6089	13449	0	0	0	0	
$E[Y D = 0, Z = 1, X]$	6194	9778	10114	13944	11752	15345	16299	18776	21010	16767	16904	12636	
$E[Y D = 0, Z = 0, X]$	8888	10980	11546	13451	12196	16084	18661	20028	19781	23669	19429	14918	
$P(D = 1 Z = 1, X)$	0.588	0.61	0.601	0.622	0.626	0.676	0.702	0.65	0.688	0.678	0.662	0.714	
$P(D = 1 Z = 0, X)$	0	0.005	0.019	0.009	0.012	0.016	0.014	0.029	0	0	0	0	

This table reports the empirical means and distributions when  $X$  is years of education.  $Y$  denotes the outcome variable which is 30-month earnings in US Dollars.  $D$  denotes the program participation and equals 1 for individuals who participated in the program.  $Z$  denotes the random assignment to treatment and equals 1 for individuals who got offered job training.

# F Additional Figures

Figure 7: Estimated conditional mean treatment responses when X is years of education



Left figure is for  $\hat{\eta}(1, X)$  and right figure is for  $\hat{\eta}(0, X)$

Figure 8: Estimated propensity score when X is years of education

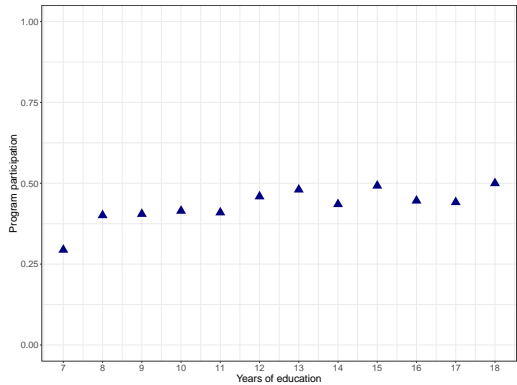
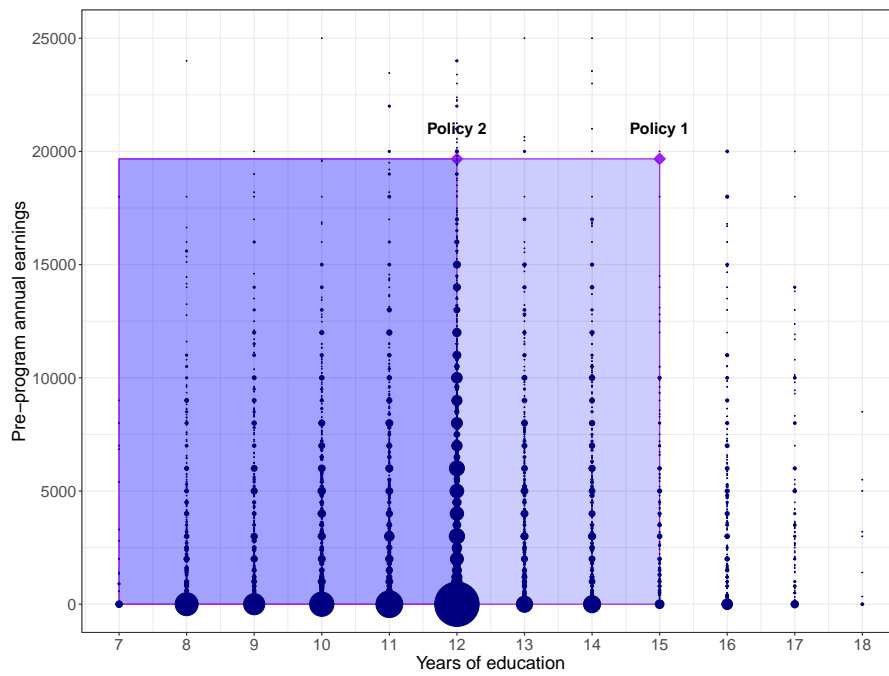
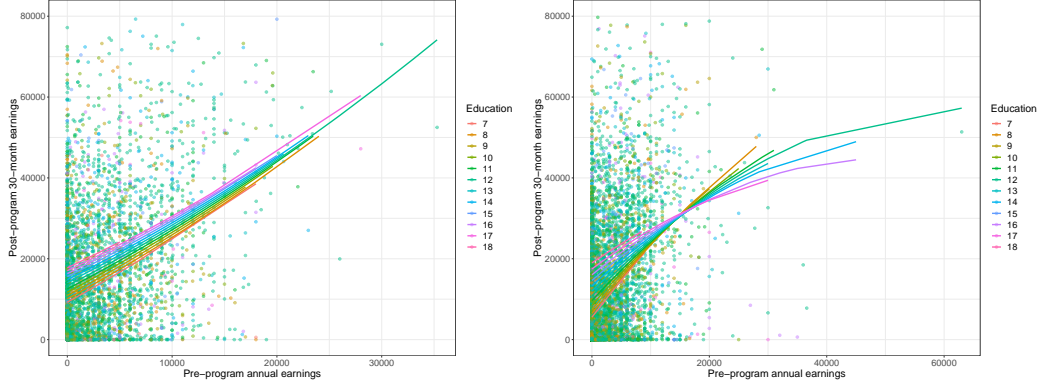


Figure 9: Hypothetical policies considered in Example 2  
in Empirical Application



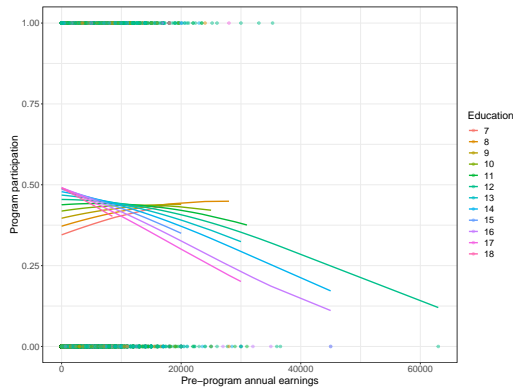
Policy 1 is  $\mathbb{1}\{\text{education} \leq 15, \text{pre-program earnings} \leq \$19,670\}$ ,  
and Policy 2 is  $\mathbb{1}\{\text{education} \leq 12, \text{pre-program earnings} \leq \$19,670\}$ .

Figure 10: Estimated conditional mean treatment response when  $X$  is years of education and pre-program annual earnings



Left figure is for  $\hat{\eta}(1, X)$  and right figure is for  $\hat{\eta}(0, X)$

Figure 11: Estimated propensity score when  $X$  is years of education and pre-program annual earnings



## References

- ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): “Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings,” *Econometrica*, 70(1), 91–117.
- ADUSUMILLI, K., F. GEIECKE, AND C. SCHILTER (2019): “Dynamically optimal treatment allocation using Reinforcement Learning,” *arXiv preprint arXiv:1904.01047*.
- ANDREWS, D. W. (1994): “Asymptotics for semiparametric econometric models via stochastic equicontinuity,” *Econometrica: Journal of the Econometric Society*, pp. 43–72.
- ANDREWS, I., T. KITAGAWA, AND A. MCCLOSKEY (2019): “Inference on winners,” Discussion paper, National Bureau of Economic Research.
- ARMSTRONG, T. B., AND S. SHEN (2015): “Inference on Optimal Treatment Assignments,” .
- ASSUNÇÃO, J., R. MCMILLAN, J. MURPHY, AND E. SOUZA-RODRIGUES (2019): “Optimal Environmental Targeting in the Amazon Rainforest,” Discussion paper, National Bureau of Economic Research.
- ATHEY, S., G. W. IMBENS, AND S. WAGER (2018): “Approximate residual balancing: debiased inference of average treatment effects in high dimensions,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4), 597–623.
- ATHEY, S., AND S. WAGER (2020): “Policy Learning with Observational Data,” .
- BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): “Sparse models and methods for optimal instruments with an application to eminent domain,” *Econometrica*, 80(6), 2369–2429.
- BELLONI, A., V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN (2017): “Program evaluation and causal inference with high-dimensional data,” *Econometrica*, 85(1), 233–298.
- BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): “Inference on treatment effects after selection among high-dimensional controls,” *The Review of Economic Studies*, 81(2), 608–650.
- BERESTEANU, A., I. MOLCHANOV, AND F. MOLINARI (2011): “Sharp identification regions in models with convex moment predictions,” *Econometrica*, 79(6), 1785–1821.
- (2012): “Partial identification using random set theory,” *Journal of Econometrics*, 166(1), 17–32.
- BERESTEANU, A., AND F. MOLINARI (2008): “Asymptotic properties for a class of partially identified models,” *Econometrica*, 76(4), 763–814.
- BHATTACHARYA, D., AND P. DUPAS (2012): “Inferring welfare maximizing treatment assignment under budget constraints,” *Journal of Econometrics*, 167(1), 168–196.

- BLOOM, H. S., L. L. ORR, S. H. BELL, G. CAVE, F. DOOLITTLE, W. LIN, AND J. M. BOS (1997): “The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act study,” *Journal of human resources*, pp. 549–576.
- BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): “Set identified linear models,” *Econometrica*, 80(3), 1129–1155.
- CHAMBERLAIN, G. (2011): “Bayesian aspects of treatment choice,” in *The Oxford Handbook of Bayesian Econometrics*.
- 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,” .
- CHERNOZHUKOV, V., J. C. ESCANCIANO, H. ICHIMURA, W. K. NEWEY, AND J. M. ROBINS (2020): “Locally Robust Semiparametric Estimation,” .
- CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, J. HAHN, AND W. NEWEY (2013): “Average and quantile effects in nonseparable panel models,” *Econometrica*, 81(2), 535–580.
- CHESHER, A., AND A. M. ROSEN (2017): “Generalized instrumental variable models,” *Econometrica*, 85(3), 959–989.
- CUI, Y., AND E. TCHETGEN TCHETGEN (2020): “A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity,” *Journal of the American Statistical Association*, pp. 1–12.
- DEHEJIA, R. H. (2005): “Program evaluation as a decision problem,” *Journal of Econometrics*, 125(1), 141–173.
- EPSTEIN, L. G., H. KAIDO, AND K. SEO (2016): “Robust confidence regions for incomplete models,” *Econometrica*, 84(5), 1799–1838.
- FARRELL, M. H. (2015): “Robust inference on average treatment effects with possibly more covariates than observations,” *Journal of Econometrics*, 189(1), 1–23.
- GALICHON, A., AND M. HENRY (2011): “Set identification in models with multiple equilibria,” *The Review of Economic Studies*, 78(4), 1264–1298.
- HAN, S. (2019): “Optimal Dynamic Treatment Regimes and Partial Welfare Ordering,” .
- HECKMAN, J. J., AND E. J. VYTLACIL (2007): “Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation,” *Handbook of econometrics*, 6, 4779–4874.
- HIRANO, K., AND J. R. PORTER (2009): “Asymptotics for statistical treatment rules,” *Econometrica*, 77(5), 1683–1701.

- ICHIMURA, H., AND W. K. NEWAY (2017): “The influence function of semiparametric estimators,” *arXiv preprint arXiv:1508.01378*.
- IMBENS, G. W., AND D. B. RUBIN (2015): *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- KAIDO, H. (2016): “A dual approach to inference for partially identified econometric models,” *Journal of econometrics*, 192(1), 269–290.
- (2017): “Asymptotically efficient estimation of weighted average derivatives with an interval censored variable,” *Econometric Theory*, 33(5), 1218–1241.
- KAIDO, H., AND A. SANTOS (2014): “Asymptotically efficient estimation of models defined by convex moment inequalities,” *Econometrica*, 82(1), 387–413.
- KAIDO, H., AND Y. ZHANG (2019): “Robust Likelihood Ratio Tests for Incomplete Economic Models,” *arXiv preprint arXiv:1910.04610*.
- KASY, M. (2014): “Using data to inform policy,” .
- (2016): “Partial identification, distributional preferences, and the welfare ranking of policies,” *Review of Economics and Statistics*, 98(1), 111–131.
- KITAGAWA, T., AND A. TETENOV (2018): “Who should be treated? empirical welfare maximization methods for treatment choice,” *Econometrica*, 86(2), 591–616.
- (2019): “Equality-minded treatment choice,” *Journal of Business & Economic Statistics*, (just-accepted), 1–40.
- KOCK, A. B., D. PREINERSTORFER, AND B. VELIYEV (2018): “Functional Sequential Treatment Allocation,” *arXiv preprint arXiv:1812.09408*.
- KOCK, A. B., AND M. THYRSGAARD (2017): “Optimal sequential treatment allocation,” .
- MANSKI, C. F. (1990): “Nonparametric bounds on treatment effects,” *The American Economic Review*, 80(2), 319–323.
- (1997): “Monotone treatment response,” *Econometrica: Journal of the Econometric Society*, pp. 1311–1334.
- (2003): *Partial identification of probability distributions*. Springer Science & Business Media.
- (2004): “Statistical treatment rules for heterogeneous populations,” *Econometrica*, 72(4), 1221–1246.
- MANSKI, C. F., AND J. V. PEPPER (2000): “Monotone instrumental variables: with an application to the returns to schooling,” *Econometrica*, 68(4), 997–1010.



- MBAKOP, E., AND M. TABORD-MEEHAN (2016): “Model Selection for Treatment Choice: Penalized Welfare Maximization,” *arXiv preprint arXiv:1609.03167*.
- MOLCHANOV, I., AND F. MOLINARI (2018): *Random Sets in Econometrics*. Cambridge University Press.
- MOLCHANOV, I. S. (2017): *Theory of Random Sets*. Springer.
- MOLINARI, F. (2019): “Econometrics with Partial Identification,” *Handbook of Econometrics*.
- NEWHEY, W. K. (1990): “Semiparametric efficiency bounds,” *Journal of applied econometrics*, 5(2), 99–135.
- (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.
- NEWHEY, W. K., AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” *Handbook of econometrics*, 4, 2111–2245.
- PONOMAREVA, M., AND E. TAMER (2011): “Misspecification in moment inequality models: Back to moment equalities?,” *The Econometrics Journal*, 14(2), 186–203.
- RAI, Y. (2018): “Statistical inference for treatment assignment policies,” *Unpublished Manuscript*.
- ROBINS, J. M., AND A. ROTNITZKY (1995): “Semiparametric efficiency in multivariate regression models with missing data,” *Journal of the American Statistical Association*, 90(429), 122–129.
- SAKAGUCHI, S. (2019): “Estimating Optimal Dynamic Treatment Assignment Rules under Intertemporal Budget Constraints,” .
- SASAKI, Y., AND T. URA (2018): “Estimation and Inference for Policy Relevant Treatment Effects,” *arXiv preprint arXiv:1805.11503*.
- STOYE, J. (2007): “Minimax regret treatment choice with incomplete data and many treatments,” *Econometric Theory*, 23(1), 190–199.
- (2009a): “Minimax regret treatment choice with finite samples,” *Journal of Econometrics*, 151(1), 70–81.
- (2009b): “Partial identification and robust treatment choice: an application to young offenders,” *Journal of Statistical Theory and Practice*, 3(1), 239–254.
- (2012): “Minimax regret treatment choice with covariates or with limited validity of experiments,” *Journal of Econometrics*, 166(1), 138–156.
- TAMER, E. (2010): “Partial identification in econometrics,” *Annu. Rev. Econ.*, 2(1), 167–195.
- TETENOV, A. (2012): “Statistical treatment choice based on asymmetric minimax regret criteria,” *Journal of Econometrics*, 166(1), 157–165.