Sharp Bounds on Treatment Effects for Policy Evaluation*

Sukjin Han
School of Economics
University of Bristol
sukjin.han@gmail.com

Shenshen Yang
Ma Yinchu School of Economics
Tianjin University
shenshenyang@tju.edu.cn

This Draft: April 3, 2022

Abstract

For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging under unobserved heterogeneity, as is well featured in the definition of the local average treatment effect (LATE). Being intrinsically local, the LATE is known to lack external validity in counterfactual environments. This paper investigates the possibility of extrapolating local treatment effects to different counterfactual settings when instrumental variables are only binary. We propose a novel framework to systematically calculate sharp nonparametric bounds on various policy-relevant treatment parameters that are defined as weighted averages of the marginal treatment effect (MTE). Our framework is flexible enough to incorporate a large menu of identifying assumptions beyond the shape restrictions on the MTE that have been considered in prior studies. We apply our method to understand the effects of medical insurance policies on the use of medical services.

JEL Numbers: C14, C32, C33, C36

Keywords: Heterogeneous treatment effects, local average treatment effects, marginal treatment effects, extrapolation, partial identification.

*The authors are grateful to Jason Abrevaya, Brendan Kline, Xun Tang, Alex Torgovitsky, Ed Vytlacil, Haiqing Xu, and participants in the Royal Economic Society 2021 Annual Conference, the 2021 North American Summer Meeting, the 2021 Asian Meeting, the 2020 North American Winter Meeting of the Econometric Society, the 2020 Texas Econometrics Camp, and the workshop at UT Austin for helpful comments and discussions.

1 Introduction

For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging in the presence of unobserved heterogeneity. This challenge is well featured in the definition of the local average treatment effect (LATE). The LATE has been one of the most popular treatment parameters used by empirical researchers since it was introduced by Imbens and Angrist (1994). It induces a straightforward linear estimation method that requires only a binary instrumental variable (IV), and yet, allows for unrestricted treatment heterogeneity. The unfortunate feature of the LATE is that, as the name suggests, the parameter is intrinsically local, recovering the average treatment effect (ATE) for a specific subgroup of population called compliers. This feature leads to two major challenges in making the LATE a reliable parameter for counterfactual policy evaluation. First, the subpopulation for which the effect is measured may not be the population of policy interest. Second, the definition of the subpopulation depends on the IV chosen, rendering the parameter even more difficult to extrapolate to new environments.

Dealing with the lack of external validity of the LATE has been an important theme in the literature. One approach in theoretical work (Angrist and Fernandez-Val (2010); Bertanha and Imbens (2019)) and empirical research (Dehejia et al. (2019); Muralidharan et al. (2019)) has been to show the similarity between complier and non-complier groups based on observables. This approach, however, cannot attend to possible unobservable discrepancies between these groups. Heckman and Vytlacil (2005) unify well-known treatment parameters by expressing them as weighted averages of what they define as the marginal treatment effect (MTE). This MTE framework has a great potential for extrapolation because a class of treatment parameters that are policy-relevant can also be generated as weighted averages of the MTE.[1] The only obstacle is that the MTE is identified via a method called local IV (Heckman and Vytlacil (1999)), which requires the continuous variation of the IV that is sometime large depending on the targeted support. This in turn reflects the intrinsic difficulty of extrapolation when available exogenous variation is only discrete. Acknowledging this nature of the challenge, previous studies in the literature have proposed imposing shape restrictions on the MTE, which is a function of the treatment-selection unobservable, while allowing for binary instruments in the framework of Heckman and Vytlacil (2005). Brinch et al. (2017) introduce shape restrictions (e.g., linearity) on the MTE functions in an attempt to identify the LATE extrapolated to different subpopulations or to test for its externality validity. In interesting recent work, Mogstad et al. (2018) propose a general partial identification frame-

[1] See Heckman (2010) for elaboration of this point.

work where bounds on various policy-relevant treatment parameters can be obtained from a set of "IV-like estimands" that are directly identified from the data and routinely obtained in empirical work. Kowalski (2020) applies an approach similar to these studies to extrapolate the results from one health insurance experiment to an external setting.

This paper continues this pursuit and investigates the possibility of extrapolating local treatment parameters to different policy settings in the MTE framework when IVs are only binary. We show how to systematically calculate sharp nonparametric bounds on various extrapolated treatment parameters for binary and continuous outcomes using instruments that are allowed to be binary. These parameters are defined as weighted averages of the MTE. Examples include the ATE, the treatment effect on the treated, the LATE for subgroups induced by new policies, and the policy-relevant treatment effect (PRTE). We also show how to place in this procedure restrictions from a large menu of identifying assumptions beyond the shape restrictions considered in earlier work.

In this paper, we make four main contributions. First, we introduce identifying assumptions that have not been used in the context of the MTE framework or the LATE extrapolation. They include assumptions that there exist exogenous variables other than IVs. One of the main messages we hope to deliver in this paper is that, given the challenge of extrapolation, additional exogenous variation can be useful to conduct informative policy evaluation. We propose two types of exogenous variables that have been used in the literature in the context of identifying the ATE: Mourifié (2015), Han and Vytlacil (2017), Vuong and Xu (2017), and Han and Lee (2019) use the first type (entering the outcome and selection equations), and Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and Balat and Han (2018) use the second type (only entering the outcome equation). We utilize these variables in this novel context of the MTE framework. Moreover, while the existing papers on the ATE exploit these variables in combination with rank similarity or rank invariance assumptions, we show that they independently have identifying power for treatment parameters, including the ATE. Of course, it may not be always easy to find such exogenous variation. But when the researcher does find it, it can be a more reliable source of identification than assumptions on unobservable quantities (e.g., shape restrictions on the MTE), as the identifying power comes from the data rather than the researcher's prior. The assumptions on extra variations can be combined with restrictions on unobservables. In this sense, we view this approach complementary to the earlier work on extrapolation mentioned above.

We also propose identifying assumptions that restrict treatment effect heterogeneity. In particular, we propose a range of uniformity assumptions that are weaker than rank similarity or rank invariance (Chernozhukov and Hansen (2005)), including a novel identifying assumption, called *rank dominance*. The direction of endogeneity can also be incorporated

in this MTE framework. This assumption is sometimes imposed in empirical work to characterize selection bias and has been shown to have identifying power for the ATE (Manski and Pepper (2000)).

Second, in order to operationalize the use of these identifying assumptions, we propose a novel framework for calculating bounds on policy-relevant treatment parameters. We introduce the distribution of the latent state of the outcome-generating process conditional on the treatment-selection unobservable. This latent conditional distribution is the key ingredient for our analysis, as both the target parameter and the distribution of the observables can be written as linear functionals of it. Because the latent distribution is a fundamental quantity in the data-generating process, it is convenient to impose identifying assumptions. Having the latent distribution as a decision variable, we can formulate infinite-dimensional linear programming (LP) that produces bounds on a targeted treatment parameter. Our approach is reminiscent of Balke and Pearl (1997) and can be viewed as its generalization to the MTE framework. Balke and Pearl (1997) introduce a LP approach to characterize bounds on the ATE with a binary outcome, treatment and instrument. The main distinction of our approach is that the latent distribution is conditioned on the selection unobservable, which makes the program infinite-dimensional, but is important for our extrapolation purpose. To make it feasible to solve the resulting infinite-dimensional program, we use a sieve-like approximation of the program and produce a finite-dimensional LP. The use of approximation is similar to Mogstad et al. (2018)'s approach, although the latter directly approximates the MTE function, which is their ingredient to relate IV-like estimands, shape restrictions, and target parameters. We also develop a method to rescale the LP to resolve computational issues that arise with a large sieve dimension.

Third, we show that our approach yields straightforward proof of the sharpness of the resulting bounds, no matter whether the outcome is discrete or continuous and whether additional identifying assumptions are imposed or not. This feature stems from the use of the latent conditional distribution in the linear programming and the convexity of the feasible set in the program. When the MTE itself is the target parameter, we distinguish between the notions of point-wise and uniform sharpness and argue why uniform sharpness is often difficult to achieve.

Fourth, as an application, we study the effects of insurance on medical service utilization by considering various counterfactual policies related to insurance coverage. The LATE for compliers and the bounds on the LATE for always-takers and never-takers reveal that possessing private insurance has the largest effect on medical visits for never-takers, i.e., those who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because the high cost might hinder people with most need

from receiving adequate medical services.

The linear programming approach to partial identification of treatment effects was pioneered by Balke and Pearl (1997) and recently gained attention in the literature; see, e.g., Mogstad et al. (2018), Torgovitsky (2019a), Machado et al. (2019), Kamat (2019), Gunsilius (2019), and Han (2019). As these papers suggest, there are many settings, including ours, where analytical derivation of bounds is cumbersome or nearly impossible due to the complexity of the problems. As concurrent work to ours, Marx (2020) considers partial identification in the MTE framework. In his paper, sharp analytical bounds are derived for treatment parameters and identifying power of rank similarity and covariates is explored. The current paper is similar to his in the sense that we exploit the conditional distribution of data (rather than the conditional mean as in Mogstad et al. (2018)) to produce sharp bounds. However, instead of analytical bounds, we provide a computational framework that enables the systematic calculation of bounds under various assumptions that have not been previously explored in this context (e.g., uniformity weaker than rank similarity and exogenous variables entering the model in specific ways).

This paper will proceed as follows. The next section introduces the main observables, maintained assumptions, and parameters of interest, and Section 3 introduces additional identifying assumptions. Section 4 defines the latent conditional probability and formulates the infinite-dimensional LP, and Section 5 introduces sieve approximation to the program. Section 6 shows how assumptions in Section 3 can easily be incorporated in the LP. So far, the analysis is given with binary Y, which is extended to the case with continuous Y in Section 7. Section 8 provides numerical illustrations, and Section 9 contains an empirical application. In the Appendix, Section A lists other examples of target parameters. Section B discusses (i) rescaling of the LP, (ii) the point-wise and uniform sharpness for the MTE bounds, (iii) the extension with continuous covariates, (iv) estimation and inference, and (v) the relationship between this paper's LP and that in Mogstad et al. (2018). All proofs are contained in Section C.

2 Assumptions and Target Parameters

Assume that we observe a discrete or continuous outcome $Y \in \mathcal{Y}$, binary treatment $D \in \{0,1\}$, and binary instrument $Z \in \{0,1\}$. We may additionally observe an exogenous variable $W \in \mathcal{W}$ and (possibly endogenous) covariates $X \in \mathcal{X}$. For the main analysis of the paper, we focus on binary Y as it is common in empirical work; we discuss the extension with continuously distributed Y in Section 7. Binary Z is also common especially in randomized experiments. Minimal variation in Z is the key challenge for extrapolation that we want to

address in this paper. We will mostly focus on binary W and discrete X for expositional simplicity. Section B.3 in the Appendix extends the framework to incorporate continuously distributed X. It is also straightforward to extend to allow for general discrete variables for (Y, Z, W).

Let $Y(d)$ be the counterfactual outcome given d and $Y(d, w)$ be the extended counterfactual outcome given (d, w), which are consistent with the observed outcome: $Y = \sum_{d\in\{0,1\}} 1\{D = d\}Y(d) = \sum_{d\in\{0,1\},w\in\mathcal{W}} 1\{D = d, W = w\}Y(d, w)$. We consider two different scenarios related to W: (a) W directly affects Y but not D and (b) W directly affects both Y and D. Accordingly, we maintain the following assumptions.

Assumption SEL. *(a)* $D = 1\{U \leq P(Z, X)\}$ *where* $P(Z, X) \equiv \Pr[D = 1|Z, X]$*; (b)* $D = 1\{U \leq P(Z, X, W)\}$ *where* $P(Z, X, W) \equiv \Pr[D = 1|Z, X, W]$.

Assumption EX. *(a)* $(Y(d, w), D(z)) \perp (Z, W)|X$*; (b)* $(Y(d, w), D(z, w)) \perp (Z, W)|X$.

We introduce W as an additional exogenous variable researchers may be equipped with in addition to the instrument Z. In the case of (a), such variables can be motivated by exogenous shocks that agents cannot fully anticipate when making treatment choices. One of the paper's goals is to show the identifying power of W even with its minimal variation. This is the first paper that formally introduces this type of variable in the MTE framework.[2] Assumption SEL imposes a selection model for D, which is important in motivating and interpreting marginal treatment effects later. This assumption is also equivalent to Imbens and Angrist (1994)'s monotonicity assumption (Vytlacil (2002)). We introduce the standard normalization that $U \sim Unif[0, 1]$ conditional on $X = x$.[3] Assumption EX imposes the exclusion restriction and conditional independence for Z.

In these assumptions, Case (a) is where W is a reversely excluded exogenous variable, which we call *reverse IV*. This type of exogenous variables was considered by Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and Balat and Han (2018). However, unlike those studies, we exploit W without rank similarity or rank invariance. In Case (b), we show that a reverse IV is not necessary, and W can be present in the selection equation. This type of exogenous variables was considered by Mourifié (2015), Han and Vytlacil (2017), Vuong and Xu (2017), and Han and Lee (2019), but again, unlike these papers, we do not necessarily assume rank similarity or rank invariance. Below, we combine the existence of W (for both

[2] Eisenhauer et al. (2015) allows variables of type (a), but only as a feature of agent's limited information and not as a source of identification.

[3] Note that for any index function $g(z, x)$ and an unobservable ε with any distribution, the selection model satisfies $D = 1\{\varepsilon \leq g(Z, X)\} = 1\{F_{\varepsilon|X}(\varepsilon|X) \leq F_{\varepsilon|X}(g(Z, X)|X)\} = 1\{U \leq P(Z, X)\}$, since $P(z, x) = \Pr[\varepsilon \leq g(z, x)|X = x] = \Pr[U \leq F_{\varepsilon|X}(g(z, x)|x)|X = x] = F_{\varepsilon|X}(g(z, x)|x)$ and $F_{\varepsilon|X}(\varepsilon|X) = U$ is uniformly distributed conditional on X.

scenarios) with assumptions that are related to or weaker than rank similarity. Another distinct feature of our approach in comparison to the prior studies is that we consider a broad class of the generalized LATEs as our target parameter, including the ATE considered in those studies. For notational simplicity, we focus on Case (a) henceforth; it is straightforward to draw analogous results for Case (b).

We aim to establish sharp bounds on various treatment parameters. Following Heckman and Vytlacil (2005), we express treatment parameters as integral equations of the MTE. The MTE is defined in our setting as

$$E[Y(1) - Y(0)|U = u, X = x],$$

where $Y(d) = Y(d, W)$. Similar to Mogstad et al. (2018), it is convenient to introduce the marginal treatment response (MTR) function

$$m_d(u, w, x) \equiv E[Y(d, w)|U = u, X = x]$$

where W does not appear as a conditioning variable due to Assumption EX(a). Now, we define the target parameter τ to be a weighted average of the MTE:

$$\tau = E[\tau_1(Z, W, X) - \tau_0(Z, W, X)], \tag{2.1}$$

where

$$\tau_d(z, w, x) = \int m_d(u, w, x)\omega_d(u, z, x)du \tag{2.2}$$

by using $F_{U|X}(u|x) = u$, and $\omega_d(u, z, x)$ is a known weight specific to the parameter of interest.[4] This definition agrees with the insight of Heckman and Vytlacil (2005). The target parameter includes a wide range of policy-relevant treatment parameters. We list a few examples of the target parameter here; other examples can be found in Table 4 in the Appendix.

Example 1. *With a Dirac delta function for a given value u as the weight, the MTE itself can be a target parameter.*

$$\tau_{MTE} = E[m_1(u, W, X) - m_0(u, W, X)]$$

Example 2. *The ATE can be a target parameter with $\omega_d(u, z, x) = 1$ for any (u, z, x).*

[4] Mogstad et al. (2018) define the weight in a slightly different way.

$$\tau_{ATE} = E\left[\int_0^1 m_1(u, W, X)du - \int_0^1 m_0(u, W, X)du\right]$$

Example 3. *The generalized LATE is also a target parameter. Suppose we are interested in the LATE for individuals lying in* $[\underline{u}, \overline{u}]$*. We assign the weight* $\omega_d(u, z, x) = \frac{1(u \in [\underline{u}, \overline{u}])}{\overline{u} - \underline{u}}$ *for any* (u, z, x) *, where the numerator excludes people outside this range and the denominator gives a weight to people within* $[\underline{u}, \overline{u}]$ *according to their fraction in the whole population. The generalized LATE is expressed as:*

$$\tau_{GLATE} = E\left[\int_0^1 m_1(u, W, X)\frac{1(u \in [\underline{u}, \overline{u}])}{\overline{u} - \underline{u}}du - \int_0^1 m_0(u, W, X)\frac{1(u \in [\underline{u}, \overline{u}])}{\overline{u} - \underline{u}}du\right]$$

Example 4. *The policy relevant treatment effect (PRTE) is a target parameter that is particularly useful for policy evaluation. It is defined as the welfare difference between two different policies. Let* Z *and* Z' *be two instrument variables under two policies and* $P(Z, X)$ *and* $P'(Z', X)$ *be propensity scores under the two policies.*

$$\begin{aligned}\tau_{PRTE} = E\Bigg[&\int_0^1 m_1(u, W, X)\frac{\Pr\left[u \leq P'(Z', X)\right] - \Pr\left[u \leq P(Z, X)\right]}{E\left[P'(Z', X)\right] - E\left[P(Z, X)\right]}du \\ &- \int_0^1 m_0(u, W, X)\frac{\Pr\left[u \leq P'(Z', X)\right] - \Pr\left[u \leq P(Z, X)\right]}{E\left[P'(Z', X)\right] - E\left[P(Z, X)\right]}du\Bigg]\end{aligned}$$

In these examples, the weights ω_0 and ω_1 can be set asymmetrically to define a broader class of parameters. All the parameters we consider in this paper can be defined conditional on X and W, although we omit them for succinctness.

Typically, a binary instrument is not sufficient in producing informative bounds on the target parameters. This is because a binary instrument has no extrapolative power for general non-compliers, e.g., always-takers and never-takers, but only identifies the effect for compliers. Prior studies have tried to overcome this challenge by imposing shape restrictions on the MTE (Cornelissen et al. (2016), Brinch et al. (2017), Mogstad et al. (2018), Kowalski (2020)), although these restrictions are not always empirically justified. Evidently, it would be useful to provide empirical researchers with a larger variety of assumptions so that it is easier to find justifiable assumptions that suit their specific examples. As such, we introduce two assumptions here. In the following section, we continue listing other identifying assumptions that can be used in our framework to tighten the bounds.

The existence of additional exogenous variables embodied in Assumptions SEL and EX

can be appealing as it can be warranted by data without invoking arbitrariness. We accompany Assumptions SEL and EX with an assumption that W and Z are relevant variables, which make the role of these variables more explicit.

Assumption R. *For given $x \in \mathcal{X}$, (i) $\Pr[Y(d,w) \neq Y(d,w')|X=x] > 0$ for some d and $w \neq w'$; (ii) either (a) $P(z,x) \neq P(z',x)$ for $z \neq z'$ and $0 < P(z,x) < 1$ for all z or (b) $P(z,x,w) \neq P(z',x,w)$ for $z \neq z'$ and $0 < P(z,x,w) < 1$ for all (z,w).*

Assumption R(i) is a relevance condition for W in determining Y. R(ii) is the standard relevance assumption for the instrument and the positivity assumption. We later show that under Assumptions SEL, EX and R, the variation of W (in addition to Z) is a useful source for extrapolation and narrowing the bounds on target parameters.

3 Additional Identifying Assumptions

In addition to Assumptions SEL, EX and R, researchers may be willing to restrict the degree of treatment heterogeneity, the direction of endogeneity, or the shape of the MTR functions. Although not necessary, such restrictions play significant roles in yielding informative bounds.

3.1 Restrictions on Treatment Heterogeneity

We present a range of restrictions on treatment heterogeneity in the order of stringency starting from the strongest.

Assumption U*. *For every $w, w' \in \mathcal{W}$ and $x \in \mathcal{X}$, either $\Pr[Y(1,w) \geq Y(0,w')|X=x] = 1$ or $\Pr[Y(1,w) \leq Y(0,w')|X=x] = 1$.*

The following assumption is weaker than Assumption U*.

Assumption U. *For every $w \in \mathcal{W}$ and $x \in \mathcal{X}$, either $\Pr[Y(1,w) \geq Y(0,w)|X=x] = 1$ or $\Pr[Y(1,w) \leq Y(0,w)|X=x] = 1$.*

When W is not available at all, this assumption can be understood with $\mathcal{W}$ being degenerate. In Assumption U*, w and w' may be the same or different, i.e., the uniformity is for all combinations of $(w,w') \in \{(0,0),(1,1),(1,0),(0,1)\}$. Therefore, Assumption U* implies Assumption U. Assumptions U and U* posit that individuals present uniformity in the sense that the treatment either weakly increases the outcome for all individuals or decreases it for all individuals. The assumptions share insights with the monotone treatment response assumption that is introduced to bound the ATE in Manski (1997) and Manski

and Pepper (2000). However, Assumptions U and U* are weaker than monotone treatment response because they do not impose the direction of monotonicity. Also, it is easy to see that Assumptions U and U* are weaker than and implied by the rank similarity and rank invariance assumptions considered in the literature (e.g., Chernozhukov and Hansen (2005), Marx (2020)). Namely, given a structural model $Y = 1[s(D, W) \geq V_D]$ with $V_D \equiv DV_1 + (1-D)V_0$, when Assumption U* is violated, then rank similarity ($F_{V_1|U} = F_{V_0|U}$) cannot hold, and thus rank invariance ($V_1 = V_0$) cannot hold. It is important to note that Assumptions U and U* still allow treatment heterogeneity in terms of X. For instance, Assumption U allows that $Y(1, w) \geq Y(0, w)$ a.s. for $X = x$ but $Y(1, w) \leq Y(0, w)$ a.s. for $X = x'$.[5]

In fact, it is possible to further weaken Assumption U. To facilitate the discussion, note that with binary Y, Assumption U can be alternatively stated as follows: for every w and given x, either $P[Y(1, w) = 0, Y(0, w) = 1|X = x] = 0$ or $P[Y(1, w) = 1, Y(0, w) = 0|X = x] = 0$. In other words, all individuals respond weakly monotonically to the treatment in the sense that there is no (strictly) negative-treatment-response type or positive-treatment-response type. This idea can be relaxed by allowing for the existence of both types in the population and instead assuming the dominance of one of the types over the other: for example, $P[Y(1, w) = 1, Y(0, w) = 0|X = x] \geq P[Y(1, w) = 0, Y(0, w) = 1|X = x]$. For general Y, such an assumption is written as follows:

Assumption U^0. *For every $w \in \mathcal{W}$ and $x \in \mathcal{X}$, either $P[Y(1, w) \geq Y(0, w)|X = x] \geq P[Y(1, w) \leq Y(0, w)|X = x]$ or $P[Y(1, w) \geq Y(0, w)|X = x] \leq P[Y(1, w) \leq Y(0, w)|X = x]$.*

Assumption U^0 is weaker than Assumption U, because when $P[Y(1, w) \leq Y(0, w)|X = x] = 0$, Assumption U^0 trivially holds.[6] Compared to Assumption U, Assumption U^0 allows further treatment heterogeneity in that positive- and negative-treatment-response types can both present in the population. Researchers may be more comfortable with this assumption than complete uniformity. This paper is the first to propose to use this restriction in the identification of treatment effects. We call assumptions of this type *rank dominance*. As shown later, the directions in Assumptions U*–U^0 can be learned from the data.

3.2 Direction of Endogeneity

In some applications, researchers are relatively confident about the direction of treatment endogeneity. The idea of imposing the direction of the selection bias as an identifying as-

[5] This idea of conditional rank preservation first appears in Han (2021) and also used in Marx (2020).

[6] One can come up with assumptions which strength is between U^0 and U. For example with binary Y, one can assume that, in addition to U^0, $P[Y(1, w) = 1, Y(0, w) = 1|X = x] \geq P[Y(1, w) = 0, Y(0, w) = 1|X = x]$ and $P[Y(1, w) = 0, Y(0, w) = 0|X = x] \geq P[Y(1, w) = 0, Y(0, w) = 1|X = x]$ hold. We do not explore these assumptions for succinctness.

sumption appears in Manski and Pepper (2000), who introduce monotone treatment selection (MTS), in addition to the monotone treatment response assumption mentioned above.

Assumption MTS. *For every* $w \in \mathcal{W}$ *and* $x \in \mathcal{X}$ $E[Y(d,w)|D=1, X=x] \geq E[Y(d,w)|D=0, X=x]$ *for* $d \in \{0,1\}$.

3.3 Shape Restrictions

It is straightforward to incorporate the shape restrictions on the MTR or MTE function introduced in the literature. They can be imposed via constraints on θ.

Assumption M. *For* $x \in \mathcal{X}$, $m_d(u,x)$ *is weakly increasing in* $u \in [0,1]$.

Assumption C. *For* $x \in \mathcal{X}$, $m_d(u,x)$ *is weakly concave in* $u \in [0,1]$.

Assumption M appears in Brinch et al. (2017) and Mogstad et al. (2018) and Assumption C appears in Mogstad et al. (2018). Another shape restriction introduced in the literature is separability: For $x \in \mathcal{X}$, $m_d(u,w,x) = m_{1d}(w,x) + m_{2d}(w,u)$ is weakly concave in $u \in [0,1]$.

4 Distribution of Latent State and Infinite-Dimensional Linear Program

Our goal is to provide a systematic framework to calculate bounds on the target parameters, which is easy to incorporate various identifying assumptions, including those introduced in Sections 2 and 3. To this end and as a crucial first step of our analysis, we define a state variable that determines a specific mapping of

$$(d,w) \mapsto y. \tag{4.1}$$

Given that d and y are binary and assuming w is also binary, there are sixteen possible maps from (d,w) onto y. Define a discrete latent variable ϵ whose value e corresponds to each possible map:

$$\epsilon \in \mathcal{E},$$

where $|\mathcal{E}| = 16$. Conveniently, let $\mathcal{E} \equiv \{1,2,...,16\}$. That is, ϵ is a decimal transformation of a binary sequence $(Y(0,0), Y(0,1), Y(1,0), Y(1,1))$, which captures all the relevant treatment effect heterogeneity. Table 1 lists all 16 maps. For the later purpose, it is helpful to explicitly

ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$
1	0	0	0	5	0	0	0	9	0	0	0	13	0	0	0
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
2	0	0	1	6	0	0	1	10	0	0	1	14	0	0	1
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
3	0	0	0	7	0	0	0	11	0	0	0	15	0	0	0
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
4	0	0	1	8	0	0	1	12	0	0	1	16	0	0	1
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1

Table 1: All Possible Maps from (d, w) to y

define the map as

$$y = g_e(d, w)$$

and write

$$Y(d, w) = g_\epsilon(d, w), \tag{4.2}$$

which implies $Y = g_\epsilon(D, W)$. It is important to note that no structure is imposed in introducing $g_e(\cdot)$ because the mapping is fully saturated. By (4.2) and Assumption SEL(a), Assumption EX(a) can be equivalently stated as $(\epsilon, U) \perp (Z, W)|X$. Still, ϵ and X can be correlated as X is allowed to be endogenous.

Now, as a key component of our LP, we define the probability mass function of ϵ conditional on (U, X): for $e \in \mathcal{E}$,

$$q(e|u, x) \equiv \Pr[\epsilon = e|U = u, X = x] = \Pr[\epsilon = e|U = u, X = x, W = w] \tag{4.3}$$

with $\sum_{e\in\mathcal{E}} q(e|u, x) = 1$ for any u, x. The quantity $q(e|u, x)$ captures endogenous treatment selection. It is shown below that this latent conditional probability is a building block for various treatment parameters and thus serves as the decision variable in the LP. The

introduction of $q(e|u,x)$ distinguishes our approach from those in Balke and Pearl (1997) and Mogstad et al. (2018). Since the probability is conditional on continuously distributed U, the simple finite-dimensional linear programming approach of Balke and Pearl (1997) is no longer applicable. Instead, we use an approximation method similar to Mogstad et al. (2018). However, Mogstad et al. (2018) uses the MTR function as a building block for treatment parameters and introduces the "IV-like" estimands as a means of funneling the information from the data. Unlike in Mogstad et al. (2018), $q(e|u,x)$ can be directly related to the distribution of data. This allows us to facilitate proving sharpness and incorporating additional identifying assumptions.

By (4.2) and (4.3), note that

$$\begin{aligned}\Pr[Y(d)=1|U=u,X=x] &= \Pr[\epsilon \in \{e \in \mathcal{E} : g_e(d,w)=1\}|U=u,X=x] \\ &= \sum_{e\in\mathcal{E}:g_e(d,w)=1} q(e|u,x).\end{aligned}$$

Therefore, the MTR can be expressed as

$$m_d(u,w,x) = \sum_{e:g_e(d,w)=1} q(e|u,x). \tag{4.4}$$

Combining (2.2) and (4.4), we have $\tau_d(z,w,x) = \sum_{e:g_e(d,w)=1} \int q(e|u,x)\omega_d(u,z,x)du$, and thus the target parameter $\tau = E[\tau_1(Z,W,X)] - E[\tau_0(Z,W,X)]$ in (2.1) can be written as

$$\tau = E\left[\sum_{e:g_e(1,W)=1} \int q(e|u,X)\omega_1(u,Z,X)du - \sum_{e:g_e(0,W)=1} \int q(e|u,X)\omega_0(u,Z,X)du\right] \tag{4.5}$$

for some q that satisfies the properties of probability.

The goal of this paper is to (at least partially) infer the target parameter τ based on the data, i.e., the distribution of (Y,D,Z,W,X). The key insight is that there are observationally equivalent $q(e|u,x)$'s that are consistent with the data, which in turn produces observationally equivalent τ's that define the identified set.

Let $p(y,d|z,w,x) \equiv \Pr[Y=y,D=d|Z=z,W=w,X=x]$ be the observed conditional probability. This data distribution imposes restrictions on $q(e|u,x)$. For instance, for $D=1$,

$$\begin{aligned}p(y,1|z,w,x) &= \Pr[Y(1,w)=y, U \le P(z,x)|Z=z,W=w,X=x] \\ &= \Pr[Y(1,w)=y, U \le P(z,x)|X=x]\end{aligned}$$

by Assumption EX(a), but

$$\Pr[Y(1,w)=y, U\leq P(z,x)|X=x] = \int_0^{P(z,x)} \Pr[Y(1,w)=y|U=u,X=x]du$$
$$= \sum_{e:g_e(1,w)=y} \int_0^{P(z,x)} q(e|u,x)du, \tag{4.6}$$

where the second equality is by $\Pr[Y(d,w)=y|U=u,X=x] = \sum_{e:g_e(d,w)=y} q(e|u,x)$.

To define the identified set for τ, we introduce some simplifying notation. Let $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ and

$$\mathcal{Q} \equiv \{q(\cdot) : \sum_{e\in\mathcal{E}} q(e|u,x) = 1 \text{ and } q(e|u,x) \geq 0 \,\forall (e,u,x)\}$$

be the class of $q(u,x)$, and let $p \equiv \{p(1,d|z,w,x)\}_{(d,z,w,x)\in\{0,1\}^2\times\mathcal{W}\times\mathcal{X}}$. Also, let $R_\tau : \mathcal{Q} \to \mathbb{R}$ and $R_0 : \mathcal{Q} \to \mathbb{R}^{d_p}$ (with d_p being the dimension of p) denote the linear operators of $q(\cdot)$ that satisfy

$$R_\tau q \equiv E\left[\sum_{e:g_e(1,W)=1} \int q(e|u,X)\omega_1^\tau(u,Z,X)du - \sum_{e:g_e(0,W)=1} \int q(e|u,X)\omega_0^\tau(u,Z,X)du\right],$$
$$R_0 q \equiv \left\{\sum_{e:g_e(d,w)=1} \int_{\mathcal{U}_{z,x}^d} q(e|u,x)du\right\}_{(d,z,w,x)\in\{0,1\}^2\times\mathcal{W}\times\mathcal{X}},$$

where $\mathcal{U}_{z,x}^d$ denotes the intervals $\mathcal{U}_{z,x}^1 \equiv [0, P(z,x)]$ and $\mathcal{U}_{z,x}^0 \equiv (P(z,x), 1]$. Then, we can characterize the baseline identified set for τ where we only impose modeling primitives. Later, we show how to characterize the identified set with additional assumptions introduced in Section 3.

Definition 4.1. *Suppose Assumptions SEL(a) and EX(a) hold. The identified set of τ is defined as*

$$\mathcal{T}^* \equiv \{\tau \in \mathbb{R} : \tau = R_\tau q \text{ for some } q \in \mathcal{Q} \text{ such that } R_0 q = p\}.$$

In what follows, we formulate the infinite-dimensional LP (∞-LP) that characterizes $\mathcal{T}^*$. This program conceptualizes sharp bounds on τ from the data and the maintained

assumptions (Assumptions SEL and EX). The upper and lower bounds on τ are defined as

$$\overline{\tau} = \sup_{q \in \mathcal{Q}} R_\tau q, \tag{∞-LP1}$$

$$\underline{\tau} = \inf_{q \in \mathcal{Q}} R_\tau q, \tag{∞-LP2}$$

subject to

$$R_0 q = p. \tag{∞-LP3}$$

Observe that the set of constraints (∞-LP3) does not include

$$\sum_{e: g_e(d,w)=0} \int_{\mathcal{U}^d_{z,x}} q(e|u,x) du = p(0,d|z,w,x) \qquad \forall (d,z,w,x) \in \{0,1\}^2 \times \mathcal{W} \times \mathcal{X}. \tag{4.7}$$

This is because we know a priori that they are redundant in the sense that they do not further restrict the *feasible set*, i.e., the set of $q(e|u,x)$'s that satisfy all the constraints ($q \in \mathcal{Q}$ and (∞-LP3)).

Lemma 4.1. *In the linear program (∞-LP1)–(∞-LP3), the feasible set defined by $q \in \mathcal{Q}$ and (∞-LP3) is identical to the feasible set defined by $q \in \mathcal{Q}$, (∞-LP3), and (4.7).*

Theorem 4.1. *Under Assumptions SEL and EX, suppose $\mathcal{T}^*$ is non-empty. Then, the bounds $[\underline{\tau}, \overline{\tau}]$ in (∞-LP1)–(∞-LP3) are sharp for the target parameter τ, i.e., $cl(\mathcal{T}^*) = [\underline{\tau}, \overline{\tau}]$, where $cl(\cdot)$ is the closure of a set.*

The result of this theorem is immediate due to the convexity of the feasible set $\{q : q \in \mathcal{Q}\} \cap \{q : R_0 q = p\}$ in the LP and the linearity of $R_\tau q$ in q, which implies that $[\underline{\tau}, \overline{\tau}]$ is convex.

5 Sieve Approximation and Finite-Dimensional Linear Programming

Although conceptually useful, the LP (∞-LP1)–(∞-LP3) is not feasible in practice because $\mathcal{Q}$ is an infinite-dimensional space. In this section, we approximate (∞-LP1)–(∞-LP3) with a finite-dimensional LP via a sieve approximation of the conditional probability $q(e|u,x)$. We use Bernstein polynomials as the sieve basis. Bernstein polynomials are useful in imposing restrictions on the original function (Joy (2000); Chen et al. (2011); Chen et al. (2017)) and therefore have been introduced in the context of linear programming (Mogstad et al. (2018); Masten and Poirier (2018); Mogstad et al. (2019)).

Consider the following sieve approximation of $q(e|u,x)$ using Bernstein polynomials of order K

$$q(e|u,x) \approx \sum_{k=1}^{K} \theta_k^{e,x} b_k(u),$$

where $b_k(u) \equiv b_{k,K}(u) \equiv \binom{K}{k} x^k (1-x)^{K-k}$ is a univariate Bernstein basis, $\theta_k^{e,x} \equiv \theta_{k,K}^{e,x} \equiv q(e|k/K,x)$ is its coefficient, and K is finite. It is important to note that x can index θ, because $q(e|u,x)$ is a saturated function of x. By the definition of the Bernstein coefficient, for any (e,x), it satisfies $q(e|u,x) \geq 0$ for all u if and only if $\theta_k^{e,x} \geq 0$ for all k. Also, $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) is approximately equivalent to $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x). To see this, first, $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) implies $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = \sum_{e\in\mathcal{E}} q(e|k/K,x) = 1$ for all (k,x). Conversely, when $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x),

$$\sum_{e\in\mathcal{E}} q(e|u,x) \approx \sum_{e\in\mathcal{E}} \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) = \sum_{k=1}^{K} b_k(u) = 1$$

by the binomial theorem (Coolidge (1949)). Motivated by this approximation, we formally define the following sieve space for $\mathcal{Q}$:

$$\mathcal{Q}_K \equiv \left\{ \left\{ \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) \right\}_{e\in\mathcal{E}} : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1 \text{ and } \theta_k^{e,x} \geq 0 \, \forall (e,k,x) \right\} \subseteq \mathcal{Q}. \tag{5.1}$$

Let $\mathcal{K} \equiv \{1,...,K\}$ and $p(z,w,x) \equiv \Pr[Z=z, W=w, X=x]$. For $q \in \mathcal{Q}_K$, by (4.5) and (5.1), the target parameter $\tau = E[\tau_1(Z,W,X)] - E[\tau_0(Z,W,X)]$ can be expressed with

$$E[\tau_d(Z,W,X)] = \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{e:g_e(d,w)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \gamma_k^d(w,x), \tag{5.2}$$

where $\gamma_k^d(w,x) \equiv \sum_{z\in\{0,1\}} p(z,w,x) \int b_k(u) \omega_d(u,z,x) du$. Also, for $q \in \mathcal{Q}_K$ and $D=1$, by (4.6), we have

$$p(y,1|z,w,x) = \sum_{e:g_e(1,w)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \delta_k^1(z,x), \tag{5.3}$$

where $\delta_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u) du$.

From (5.2) and (5.3), we can expect that a finite-dimensional LP can be obtained with

respect to $\theta_k^{e,x}$. Let $\theta \equiv \{\theta_k^{e,x}\}_{(e,k,x)\in\mathcal{E}\times\mathcal{K}\times\mathcal{X}}$ and let

$$\Theta_K \equiv \left\{ \theta : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1 \text{ and } \theta_k^{e,x} \geq 0 \, \forall (e,k,x) \in \mathcal{E}\times\mathcal{K}\times\mathcal{X} \right\}.$$

Then, we can formulate the following finite-dimensional LP that corresponds to the ∞-LP in (∞-LP1)–(∞-LP3):

$$\overline{\tau}_K = \max_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \left\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \right\} \tag{LP1}$$

$$\underline{\tau}_K = \min_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \left\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \right\} \tag{LP2}$$

subject to

$$\sum_{e:g_e(d,w)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^d(z,x) = p(1,d|z,w,x) \qquad \forall (d,z,w,x) \in \{0,1\}^2\times\mathcal{W}\times\mathcal{X}. \tag{LP3}$$

One of the advantages of LP is that it is computationally very easy to solve using standard algorithms, such as the simplex algorithm. Conditional on x, assuming binary W and setting $K = 50$, we have $\dim(\theta) = 816$, and it takes only around 13 seconds to calculate $\overline{\tau}_K$ and $\underline{\tau}_K$ with moderate computing power. The increase in the support of $\mathcal{W}$ (and thus the number of maps (4.1)) only linearly increases the computation time.

The important remaining question is how to choose K in practice. We discuss this issue in Section 8. Finally, it is worth noting that, extending Proposition 4 in Mogstad et al. (2018), we may exactly calculate $\overline{\tau}$ and $\underline{\tau}$ (i.e., $\overline{\tau} = \overline{\tau}_K$ and $\underline{\tau} = \underline{\tau}_K$) under the assumptions that (i) the weight function $\omega_d(u,z,x)$ is piece-wise constant in u and (ii) the constant spline that provides the best mean squared error approximation of $q(e|u,x)$ satisfies all the maintained assumptions (possibly including the identifying assumptions introduced later) that $q(e|u,x)$ itself satisfies; see Mogstad et al. (2018) for details.

6 Incorporating Identifying Assumptions

In this section, we show how identifying assumptions introduced in Section 3 can be easily translated into assumptions on the mapping defined in (4.1). This allows us to incorporate the additional assumptions in the formulation of the LP, so that one can obtain more informative bounds.

Before proceeding, we revisit Assumptions SEL, EX, and R in the context of the LP.

First, we formally show that the existence and relevance of W (as well as Z) embodied in Assumptions SEL, EX, and R can be a useful source in narrowing the bounds.

Lemma 6.1. *Under Assumptions SEL, EX, and R, the variation of Z and W respectively poses non-redundant constraints on $q \in \mathcal{Q}$ in (∞-LP1)–(∞-LP3) and analogously $\theta \in \Theta_K$ in (LP1)–(LP3).*

Heuristically, the improvement occurs because, with R(i), the constraint matrix (i.e., the matrix multiplied to the vector θ in (LP3)) has greater rank with the variation of W than without. See the proof of the theorem for a formal argument. Note that non-redundant constraints on θ do not always guarantee an improvement of the bounds in (LP1)–(LP3), because these constraints may still be non-binding. Nevertheless, non-redundancy is a necessary condition for the improvement. In comparison, it is unclear how W can pose non-redundant constraints in Mogstad et al. (2018)'s framework as such variables will be subsumed in IV-like estimands as covariates.

We now show how to incorporate Assumptions U*, U, U^0, MTS, M, and C as additional equality and inequality restrictions in the LP: Given the LP (∞-LP1)–(∞-LP3), identifying assumptions can be imposed by appending

$$R_1 q = a_1, \tag{∞-LP4}$$

$$R_2 q \leq a_2, \tag{∞-LP5}$$

where R_1 and R_2 are linear operators on $\mathcal{Q}$ that correspond to equality and inequality constraints, respectively, and a_1 and a_2 are some vectors in Euclidean space. Then, analogous constraints on θ can be added to the finite-dimensional LP (LP1)–(LP3). When an assumption violates the true data-generating process, then the identified set will be empty. This corresponds to the situation where the LP does not have a feasible solution. When we reflect sampling errors, this corresponds to the case where the confidence set is empty.[7]

Assumptions U and U* are imposed in the LP by "deactivating" relevant maps. For example, suppose $Y(1,w) \geq Y(0,w)$ almost surely for all $w \in \{0,1\}$ under Assumption U. This assumption can be imposed as equality constraints (∞-LP4), i.e., in the form of

[7] In order to verify whether the identified set is empty, we need to check whether the feasible set of θ is empty. An efficient way to do this is to identify vertices of the feasible polytope, if any. This process is no simpler than the simplex algorithm that we use to solve the LP. Therefore, we recommend that one first solves the LP and check if infeasibility is reported.

$R_1 q = a_1$, using the labeling of Table 1: Suppressing x for simplicity,

$$q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,$$
$$q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0,$$

respectively, corresponding for $w = 1$ and $w = 0$. Therefore, the corresponding $\theta_k^e = 0$. Then, the effective dimension of θ will be reduced in (LP1)–(LP3) and thus yields narrower bounds. As another example, suppose the following holds almost surely under Assumption U*: $Y(1,1) \geq Y(0,0)$, $Y(1,0) \leq Y(0,1)$, $Y(1,1) \geq Y(0,1)$, and $Y(1,0) \geq Y(0,0)$. These inequalities respectively imply

$$q(2|u) = q(4|u) = q(6|u) = q(8|u) = 0,$$
$$q(5|u) = q(6|u) = q(13|u) = q(14|u) = 0,$$
$$q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,$$
$$q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0.$$

Recall the discussion that, in Assumption U (Assumption U*), the direction of monotonicity is allowed to be different for different w ((w, w') pairs). This direction will be identified from the data. Specifically, the direction can be automatically determined from the LP by inspecting whether the LP has a feasible solution; when wrong maps are removed, there is no feasible solution. Note that this result holds regardless of the existence of W. A similar argument applies to Assumption U^0. It is easy to see that the direction of the monotonicity coincides with the sign of the ATE. Previous work has discussed the role of the rank similarity assumption on determining the sign of the ATE (Bhattacharya et al. (2008); Shaikh and Vytlacil (2011); Han (2019)), and the result above shows that Assumptions U and U* play a similar role in the linear programming approach. Finally, suppose $P[Y(1,w) \geq Y(0,w)] \geq P[Y(1,w) \leq Y(0,w)]$ for all $w \in \{0,1\}$ under Assumption U^0. Then, we can generate the following inequality restrictions:

$$q(5|u) + q(7|u) + q(13|u) + q(15|u) \geq q(2|u) + q(4|u) + q(10|u) + q(12|u),$$
$$q(3|u) + q(4|u) + q(6|u) + q(7|u) \geq q(9|u) + q(10|u) + q(13|u) + q(14|u).$$

Note that, under Mogstad et al. (2018)'s framework, some implications of Assumptions U*–U^0 (but not these assumptions directly) may be imposed via the MTR function. In that case however, for example, Assumptions U and U^0 cannot play distinctive roles.[8]

[8]Consider $P[Y(1) \geq Y(0)] = 1$, which is consistent with Assumption U (suppressing (W, X)). This

Now consider Assumption MTS. This assumption can be imposed in the form of $R_2 q \leq a_2$. To see this, Assumption MTS is equivalent to

$$\sum_{e:g_e(d)=1} E\left[\int_{P(Z,X)}^{1} q(e|u)du - \int_{0}^{P(Z,X)} q(e|u)du \middle| X = x\right] \leq 0$$

for all $d, x \in \{0,1\} \times \mathcal{X}$. As is clear from this expression, Assumption MTS imposes restrictions on the joint distribution of (ϵ, U).

Finally, consider Assumptions M and C. These assumptions can be imposed as inequality constraints (∞-LP4), i.e., in the form of $R_2 q \leq a_2$. For implications on the finite-dimensional LP (LP1)–(LP3), recall that for $q \in \mathcal{Q}_K$, the MTR satisfies

$$m_d(u,x) = \sum_{e:g_e(d)=1} q(e|u,x) = \sum_{k\in\mathcal{K}} \sum_{e:g_e(d)=1} \theta_k^{e,x} b_k(u).$$

According to the property of the Bernstein polynomial, Assumption M implies that $\sum_{e:g_e(d)=1} \theta_k^{e,x}$ is weakly increasing in k, i.e.,

$$\sum_{e:g_e(d)=1} \theta_1^{e,x} \leq \sum_{e:g_e(d)=1} \theta_2^{e,x} \leq \cdots \leq \sum_{e:g_e(d)=1} \theta_K^{e,x}.$$

Assumption C implies that

$$\sum_{e:g_e(d)=1} \theta_k^{e,x} - \sum_{e:g_e(d)=1} 2\theta_{k+1}^{e,x} + \sum_{e:g_e(d)=1} \theta_{k+2}^{e,x} \leq 0 \qquad \text{for } k = 0, ..., K-2.$$

One can obtain analogous assumptions and their implications in the presence of W.

7 Extension: Continuous Y

7.1 Identified Set and Infinite-Dimensional Linear Programming

The analogous approach of LP can be applied to the case of continuous outcome variable. We consider the continuous outcome with support $\mathcal{Y} = [0,1]$ without loss of generality. As a key component of our LP, we define the following conditional distribution:

$$q_w(y_1, y_0|u,x) \equiv \Pr\left[Y(1,w) \leq y_1, Y(0,w) \leq y_0 | U = u, X = x\right].$$

implies that $m_1(u) \geq m_0(u)$, which then can be imposed as a restriction in Mogstad et al. (2018). However, $P[Y(1) \geq Y(0)] \geq P[Y(1) \leq Y(0)]$, which is consistent with Assumption U^0, also implies $m_1(u) \geq m_0(u)$.

First, we show how the data distribution imposes restrictions on $q_w(y_1, y_0|u, x)$. From the data, we observe

$$\pi(y, d|z, w, x) \equiv \Pr\left[Y \leq y, D = d|Z = z, W = w, X = x\right]$$

for all (y, d, z, w, x). Then, for example, consider the case with $d = 1$. The conditional distribution can be written as

$$\begin{aligned}
\pi(y, 1|z, w, x) &\equiv \Pr\left[Y \leq y, D = 1|Z = z, W = w, X = x\right] \\
&= \Pr\left[Y(1, w) \leq y, D(z, x) = 1|Z = z, W = w, X = x\right] \\
&= \Pr\left[Y(1, w) \leq y, U \leq P(z, x)|X = x\right] \\
&= \int_0^{P(z,x)} \Pr\left[Y(1, w) \leq y|U = u, X = x\right] du \\
&= \int_0^{P(z,x)} \Pr\left[Y(1, w) \leq y, Y(0, w) \leq 1|U = u, X = x\right] du \\
&\equiv \int_0^{P(z,x)} q_w(y, 1|u, x) du,
\end{aligned}$$

where the third equality follows from Assumption EX(a).

Similarly for the target parameters, the MTR function can be expressed as follows. For example, for $D = 0$,

$$\begin{aligned}
m_0(u, w, x) &= E\left[Y(0, w)|U = u, X = x\right] \\
&= \int_0^1 (1 - \Pr\left[Y(0, w) \leq y|U = u, X = x\right]) dy \\
&\equiv 1 - \int_0^1 q_w(1, y|u, x) dy.
\end{aligned}$$

We now define the identified set of the target parameters. Let $q^\dagger(\cdot) \equiv \{q_w(\cdot)\}_{w \in \mathcal{W}}$ be the vector of $q_w(\cdot)$'s. We introduce the class of $q^\dagger(\cdot)$ to be

$$\begin{aligned}
\mathcal{Q}^\dagger \equiv \big\{ & q^\dagger(\cdot) : 0 \leq q_w(\cdot, \cdot|u, x) du \leq 1, q_w(1, 1|u, x) = 1, q_w(0, 0|u, x) = 0, \\
& q_w(0, 1|u, x) = 0, q_w(1, 0|u, x) = 0, q_w(\cdot, \cdot|u, x) \text{ is increasing in its arguments } \forall (u, x) \big\}.
\end{aligned}$$

Define the CDF vector

$$\begin{aligned}\pi(y) &\equiv \{\pi(y,d|z,w,x)\}_{(d,z,w,x)\in\{0,1\}^2\times\mathcal{W}\times\mathcal{X}} \\ &\equiv \{(\pi(y,0|z,w,x),\pi(y,1|z,w,x))'\}_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}}\end{aligned}$$

and the linear operators $R_\tau^\dagger : \mathcal{Q}^\dagger \to \mathbb{R}$ and $R_0^\dagger : \mathcal{Q}^\dagger \to \mathbb{R}^{d_\pi}$ (with d_π being the dimension of π) of $q^\dagger(\cdot)$ that satisfy:

$$\begin{aligned}R_\tau^\dagger q^\dagger \equiv E\Bigg[&\int\left(1-\int_0^1 q_W(y,1|u,x)dy\right)\omega_1^\tau(u,Z,X)du \\ &-\int\left(1-\int_0^1 q_W(1,y|u,x)dy\right)\omega_0^\tau(u,Z,X)du\Bigg], \\ R_0^\dagger q^\dagger \equiv &\left\{\begin{pmatrix}\int_{\mathcal{U}_{z,x}^0} q_w(1,y|u,x)du \\ \int_{\mathcal{U}_{z,x}^1} q_w(y,1|u,x)du\end{pmatrix}\right\}_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}},\end{aligned}$$

where the expectation is taken over (W,Z,X) and $\mathcal{U}_{z,x}^1 = [0,P(z,x)]$ and $\mathcal{U}_{z,x}^0 = (P(z,x),1]$.

Definition 7.1. *The identified set of τ is defined as*

$$\mathcal{T}^* \equiv \{\tau\in\mathbb{R} : \tau = R_\tau^\dagger q^\dagger \text{ for some } q^\dagger\in\mathcal{Q}^\dagger \text{ such that } (R_0^\dagger q^\dagger)(y) = \pi(y) \text{ for all } y\in\mathcal{Y}\}.$$

Then the ∞-LP is formulated as:

$$\overline{\tau} = \sup_{q^\dagger\in\mathcal{Q}^\dagger} R_\tau^\dagger q^\dagger \tag{7.1}$$

$$\underline{\tau} = \inf_{q^\dagger\in\mathcal{Q}^\dagger} R_\tau^\dagger q^\dagger \tag{7.2}$$

subject to

$$(R_0^\dagger q^\dagger)(y) = \pi(y) \qquad \text{for all } y\in\mathcal{Y}. \tag{7.3}$$

Note that the LP is infinite dimensional not only because of $q^\dagger$ but also (7.3), which consists of a continuum of constraints.

7.2 Finite-Dimensional Linear Programming

Analogous to Section 5, we approximate the unknown function $q_w(\cdot)$ using multivariate bernstein polynomials:

$$q_w(y_1,y_0|u,x) \approx \sum_{\boldsymbol{k}=1}^K \theta_{\boldsymbol{k}}^{w,x} b_{\boldsymbol{k}}(y_1,y_0,u),$$

where $b_{\boldsymbol{k}}(y_1, y_0, u) \equiv b_{\boldsymbol{k},K}(y_1, y_0, u)$ is a trivariate Bernstein polynomials with $\boldsymbol{k} \equiv (k_1, k_0, k_u)$ and its coefficient $\theta^{w,x}_{\boldsymbol{k}} \equiv \theta^{w,x}_{\boldsymbol{k},K} \equiv q_w(k_1/K, k_0/K|k_u/K, x)$. Note that $\sum_{\boldsymbol{k}=1}^{K}$ stands for $\sum_{k_1,k_0,k_u=1}^{K}$. Then the data restriction can be written as a linear combination of the unknown parameters $\{\theta^{w,x}_{\boldsymbol{k}}\}_{(\boldsymbol{k},w,x)\in\mathcal{K}^3\times\mathcal{W}\times\mathcal{X}}$. For example,

$$\begin{aligned}\pi(y,1|z,w,x) &= \int_0^{P(z,x)} q_w(y,1|u,x)du \\ &= \sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} \int_0^{P(z,x)} b_{\boldsymbol{k}}(y,1,u)du \\ &\equiv \sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} \sigma^1_{\boldsymbol{k}}(y,z,x), \end{aligned} \tag{7.4}$$

where $\sigma^d_{\boldsymbol{k}}(y,z,x) \equiv \int_{\mathcal{U}^d_{z,x}} b_{\boldsymbol{k}}((1-d)+dy, d+(1-d)y, u)du$. Similarly, the target parameter can be written as, for example,

$$\begin{aligned} E\left[\tau_0(Z,W,X)\right] &= \sum_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}} p(z,w,x) \int \left(1 - \int_0^1 q_w(1,y|u,x)dy\right) \omega^\tau_d(u,z,x)du \\ &= \sum_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}} p(z,w,x) \int \omega^\tau_d(u,z,x)du \\ &\quad - \sum_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}} p(z,w,x) \int \left(\int_0^1 \sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} b_{\boldsymbol{k}}(1,y,u)dy\right) \omega^\tau_d(u,z,x)du \\ &= \sum_{(z,w,x)\in\{0,1\}\times\mathcal{W}\times\mathcal{X}} p(z,w,x) \int \omega^\tau_d(u,z,x)du \\ &\quad - \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} \sum_{z\in\{0,1\}} p(z,w,x) \int\int_0^1 b_{\boldsymbol{k}}(1,y,u)\omega^\tau_d(u,z,x)dydu \\ &\equiv c_d - \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} \zeta^0_{\boldsymbol{k}}(w,x), \end{aligned} \tag{7.5}$$

where $\zeta^d_{\boldsymbol{k}}(w,x) \equiv \sum_{z\in\{0,1\}} p(z,w,x) \int\int_0^1 b_{\boldsymbol{k}}((1-d)+dy, d+(1-d)y, u)\omega^\tau_d(u,z,x)dydu$.

To address the challenge that the constraint (7.4) is indexed by a continuous y, we proceed as follows. Note that, for any measurable function $h : \mathcal{Y} \to \mathbb{R}$, $E|h(Y)| = 0$ if and only if $h(y) = 0$ almost everywhere in $\mathcal{Y}$. Therefore, the constraint (with general d) can be replaced by:

$$E\left|\sum_{\boldsymbol{k}=1}^{K} \theta^{w,x}_{\boldsymbol{k}} \sigma^d_{\boldsymbol{k}}(Y,z,x) - \pi(Y,d|z,w,x)\right| = 0.$$

Now, redefine $\theta \equiv \{\theta_{\boldsymbol{k}}^{w,x}\}_{(w,x,\boldsymbol{k})\in\mathcal{W}\times\mathcal{X}\times\mathcal{K}^3}$ and

$$\begin{aligned}\Theta_K \equiv \{&\theta : 0 \leq \theta_{\boldsymbol{k}}^{w,x} \leq 1 \ \forall(w,x,\boldsymbol{k}),\\ &\theta_{K,K,k_u}^{w,x} = 1, \theta_{1,1,k_u}^{w,x} = 0, \theta_{1,K,k_u}^{w,x} = 0, \theta_{K,1,k_u}^{w,x} = 0 \ \forall(w,x,k_u),\\ &\theta_{1,k_{y_1},k_u}^{w,x} \leq \cdots \leq \theta_{K,k_{y_1},k_u}^{w,x} \ \forall(w,x,k_{y_1},k_u),\\ &\theta_{k_{y_0},1,k_u}^{w,x} \leq \cdots \leq \theta_{k_{y_0},K,k_u}^{w,x} \ \forall(w,x,k_{y_0},k_u)\Big\}.\end{aligned}$$

Then, the LP can be formulated as

$$\overline{\tau}_K = \min_{\theta\in\Theta_K} \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{\boldsymbol{k}=1}^{K} \theta_{\boldsymbol{k}}^{w,x} \left\{-\zeta_{\boldsymbol{k}}^1(w,x) + \zeta_{\boldsymbol{k}}^0(w,x)\right\} \tag{7.6}$$

$$\overline{\tau}_K = \max_{\theta\in\Theta_K} \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{\boldsymbol{k}=1}^{K} \theta_{\boldsymbol{k}}^{w,x} \left\{-\zeta_{\boldsymbol{k}}^1(w,x) + \zeta_{\boldsymbol{k}}^0(w,x)\right\} \tag{7.7}$$

subject to

$$E\left|\sum_{\boldsymbol{k}=1}^{K} \theta_{\boldsymbol{k}}^{w,x}\sigma_{\boldsymbol{k}}^d(Y,z,x) - \pi(Y,d|z,w,x)\right| = 0, \forall(y,d,z,w,x) \in \mathcal{Y}\times\{0,1\}^2\times\mathcal{W}\times\mathcal{X}. \tag{7.8}$$

8 Simulation

This section provides numerical results to illustrate our theoretical framework and to show the role of different identifying assumptions in improving bounds on the target parameters. For target parameters, we consider the ATE and the LATEs for always-takers (LATE-AT), never-takers (LATE-NT), and compliers (LATE-C). We calculate the bounds on them based only on the information from the data and then show how additional assumptions on the existence of additional exogenous variables, uniformity, and shape restrictions tighten the bounds.

8.1 Data-Generating Process

We generate the observables (Y, D, Z, X, W) from the following data-generating process (DGP). We assume that W is a reverse IV, i.e., we maintain Assumptions SEL(a) and EX(a). We allow covariate X to be endogenous. All the variables are set to be binary with $\Pr[Z=1] = 0.5$, $\Pr[X=1] = 0.6$ and $\Pr[W=1] = 0.4$. The treatment D is determined by Z and X through the threshold crossing model specified in Assumption SEL(a), where the propensity scores $P(z,x)$ are specified as follows: $P(0,0) = 0.1$, $P(1,0) = 0.4$, $P(0,1) = 0.4$,

and $P(1,1) = 0.7$. The outcome Y is generated from (D, X, W) through $Y = DY_1 + (1-D)Y_0$ where

$$Y_d = 1\left[m_d(U, X, W) \geq \epsilon\right] \tag{8.1}$$

and the MTR functions are defined as

$$\begin{aligned}
m_0(u,0,0) &= 0.002b_0^4(u) + 0.008b_1^4(u) + 0.014b_2^4(u) + 0.02b_3^4(u) + 0.021b_4^4(u),\\
m_1(u,0,0) &= 0.012b_0^4(u) + 0.048b_1^4(u) + 0.084b_2^4(u) + 0.12b_3^4(u) + 0.121b_4^4(u),\\
m_0(u,0,1) &= 0.034b_0^4(u) + 0.528b_1^4(u) + 0.724b_2^4(u) + 0.84b_3^4(u) + 0.861b_4^4(u),\\
m_1(u,0,1) &= 0.999b(u) + 0.999b_1^4(u) + 0.999b_2^4(u) + 0.999b_3^4(u) + 0.999b_4^4(u),\\
m_0(u,1,0) &= 0b_0^4(u) + 0.006b_1^4(u) + 0.012b_2^4(u) + 0.018b_3^4(u) + 0.019b_4^4(u),\\
m_1(u,1,0) &= 0b_0^4(u) + 0.036b_1^4(u) + 0.072b_2^4(u) + 0.108b_3^4(u) + 0.109b_4^4(u),\\
m_0(u,1,1) &= 0.25b_0^4(u) + 0.586b_1^4(u) + 0.822b_2^4(u) + 0.908b_3^4(u) + 0.919b_4^4(u),\\
m_1(u,1,1) &= 0.999b_0^4(u) + 0.999b_1^4(u) + 0.999b_2^4(u) + 0.999b_3^4(u) + 0.999b_4^4(u),
\end{aligned}$$

where b_k^K stands for the k-th basis function in the Bernstein approximation of degree K. These MTR functions are chosen to be consistent with Assumptions M and C, i.e., to be positively monotone and weakly concave in u for all $(d, x, w) \in \{0,1\}^3$. Also, the DGP in (8.1) satisfies Assumption U* because ϵ does not depend on $d = 0, 1$ and the MTR functions satisfy $m_1(u, x, w) > m_0(u, x, w)$ for all $(d, x, w) \in \{0,1\}^3$. Therefore, the DGP also satisfies Assumptions U and U^0. Following the second example in Section 3.1, the DGP satisfies the following uniform order for the counterfactual outcomes $Y(d, w)$: $Y(1,1) \geq Y(0,1) \geq Y(1,0) \geq Y(0,0)$ a.s. We generate a sample containing 1,000,000 observations and choose $K = 50$. We choose the large sample size to mimic the population. Our choice of K is discussed below. The number of unknown parameters θ in the linear programming is equal to $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1)$.

8.2 Bounds on Target Parameters under Different Assumptions

8.2.1 ATE

Figure 1 and 2 contain the bounds on the ATE under different assumptions. The true ATE value is 0.17, depicted as the solid red line in the figure. From 1, the worst-case bounds on the ATE with no additional assumptions (and without using variation from W) are $[-0.23, 0.47]$. Since the mappings do not involve W, we have $|\mathcal{E}| = 4$, and the linear programming is solved with $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 4 \times 2 \times 51 = 408$.

For comparison, we calculate the bounds that incorporate the existence of W. We build

up the target parameters with mappings involving W and use data distribution conditional on $W = 0$ and $W = 1$ as the constraints. Using constraints conditional on different values of W allows us to fully exploit the variations from W; see (LP3). With binary W, we have $|\mathcal{E}| = 16$, which gives $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 16 \times 2 \times 51 = 1,632$. The resulting bounds are depicted in the dotted greenish-blue line. When the variation from W is exploited, the bounds on the ATE are $[-0.13, 0.44]$, which is narrower than without using W. This result is consistent with our theoretical finding presented in Lemma 6.1 that W can help tighten the bounds as long as it is a relevant variable. Nonetheless, these worst-case bounds are not that informative, e.g., they do not determine the sign of the ATE.

Next, we impose Assumption U^0 without W and with W. Assumption U and U^0 give the same bounds in our exercise, therefore, we use the weaker version and present the results. Under Assumption U^0, the bounds on the ATE are tightened as we incorporate extra inequality constraints according to the direction of monotonicity. As mentioned in Section 3.1, the direction of monotonicity in Assumption U^0 is determined by the LPs. We solve the LPs with different directions imposed, then choose the one with a feasible solution. This means that the corresponding direction of monotonicity is consistent with the DGP. Under Assumption U^0, we obtain a bound $[0.06, 0.47]$, which is narrower comparing with the worst-case bound. With W, under Assumption U^0, the bounds become $[0.06, 0.44]$. In Figure 1, these bounds under Assumptions U^0 without and with W are depicted as violet and green dashed lines, respectively. Both sets of bounds identify the sign of the ATE, consistent with the theoretical discussion. The improvement is majorly on the lower bounds and the upper bounds coincide with the corresponding worst-case upper bound without and with W. These improvements are from the ability to identify the sign under the uniformity assumptions.

Next, we impose the shape restrictions (Assumptions M and C). As discussed in Section 3.3, these assumptions can be easily incorporated in the linear programming by directly imposing inequality constraints on θ. Under these assumptions (and the existence of W), the bounds on the ATE shrink to $[0.13, 0.21]$, which is displayed with the pink line in Figure 1. We find that shape restrictions are powerful assumptions and yield narrower bounds compared to those with uniformity assumptions. They function differently in the linear programming: unlike the uniformity assumption, which maintains the ranking of individuals across counterfactual groups, shape restrictions directly control the MTR functions.

Figure 2 presents the results under Assumption U^0, versus under Assumption U^* with existence of W. Under Assumption U^*, the bounds become $[0.06, 0.4]$. While their lower bounds coincide, Assumption U^* yields a lower upper bound compared to Assumption U^0.

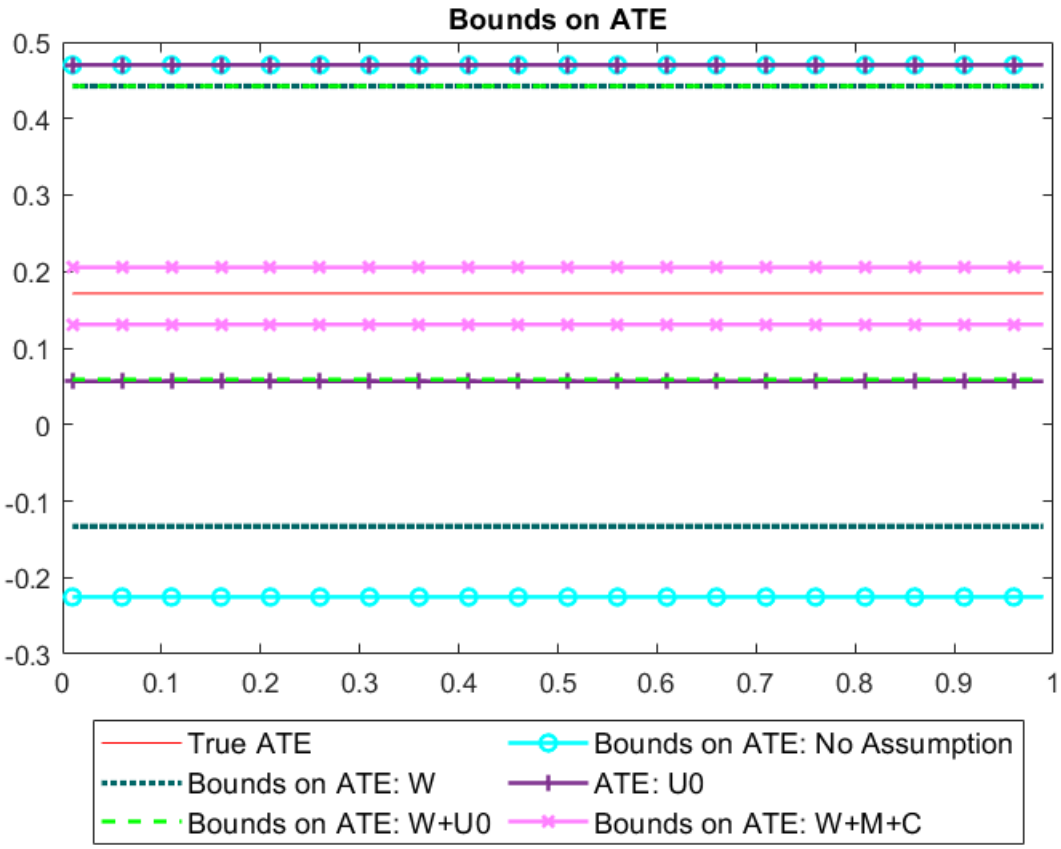


Figure 1: Bounds on the ATE under Different Assumptions

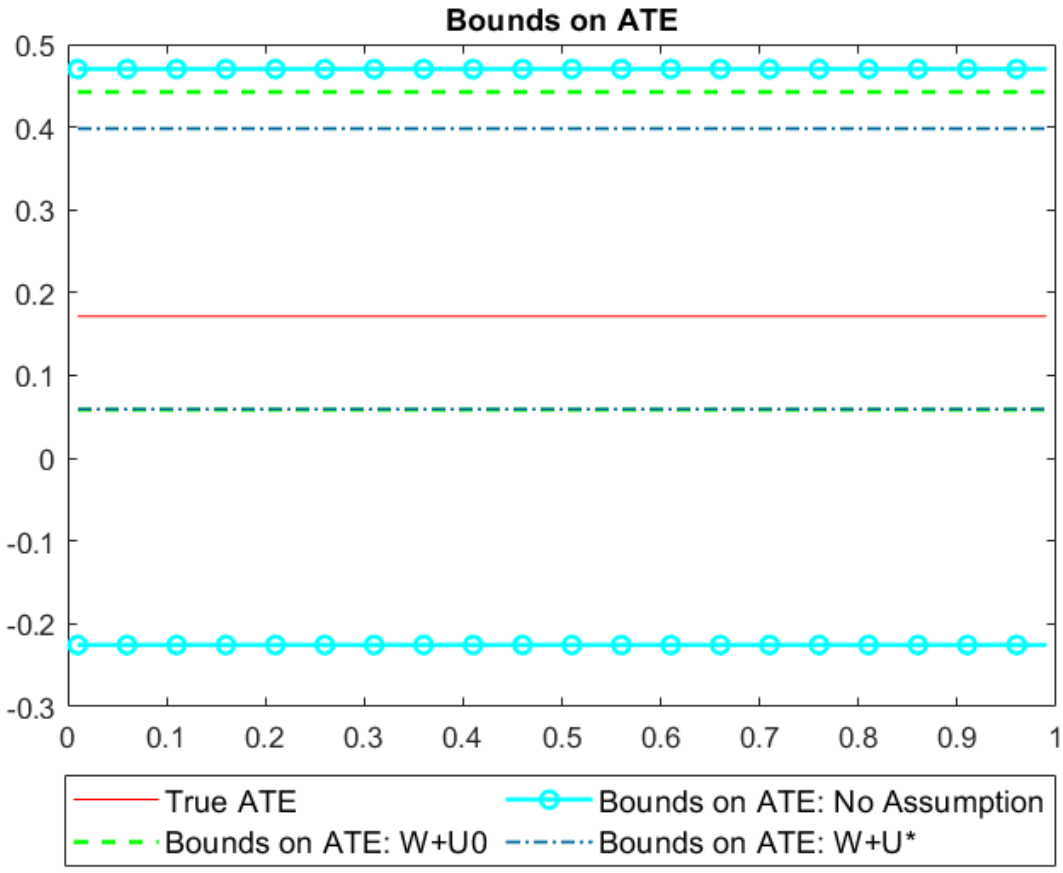


Figure 2: Bounds on the ATE under Different Assumptions

8.2.2 Generalized LATEs

Next, we construct bounds on the generalized LATEs. The original definition of the LATE is the ATE for compliers (C). Researchers may also have interests in other local treatment effects. We consider two other parameters—LATEs for always-takers (AT) and never-takers (NT). Figure 3 and 4 display the bounds on the LATE-AT, LATE-C, and LATE-NT under different assumptions. This analysis is analogous to that with the ATE. Since the covariate X affects the decision of compliance, to avoid confusion in the definition of the compliance groups, we instead establish bounds on the LATEs conditional on X. We draw the conditional MTE functions with solid red lines in both panels as a reference.

The feature that there exists no defiers in the DGP is known. When there is no defier, the LATE-C is point identified, which has an analytical expression of the two-stage least squares estimand. Therefore, even when we add the tuning parameters, the estimates remain very close to the true values throughout. And when we do not need tuning parameters to adjust the numerical errors or when the tuning parameters are very small, the linear programming yields point estimates as shown in Figure 3. The true LATE-Cs conditional on $X = 0$ and $X = 1$ are equal to 0.23 and 0.13, respectively.

The true values of the conditional LATE-AT and the LATE-NT are 0.29 and 0.14 when $X = 0$ and 0.22 and 0.09 when $X = 1$. First, as before, we consider the worst-case bounds where the existence of W is ignored versus where W is taken into account. Without W, we get the bounds $[-0.52, 0.42]$ and $[-0.27, 0.73]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.57, 0.42]$ and $[-0.39, 0.61]$ conditional on $X = 1$; with W, we get the bounds $[-0.06, 0.4]$ and $[-0.19, 0.66]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.45, 0.41]$ and $[-0.22, 0.57]$ conditional on $X = 1$. Incorporating information from W helps to improve both the upper bound and the lower bound. We then apply Assumption U^0, M and C. The bounds on the LATE-AT and the LATE-NT turn to $[0.29, 0.4]$ and $[0.13, 0.21]$ conditional on $X = 0$, and $[0.1, 0.3]$ and $[0.05, 0.13]$ conditional on $X = 1$.

From Figure 4, under the Assumption U^*, the bounds shrink to $[0, 0.4]$ and $[0, 0.53]$ conditional on $X = 0$, and $[0, 0.4]$ and $[0, 0.56]$ conditional on $X = 1$, comparing with the bounds under Assumption U^0. The improvement is from complete order of 16 mapping types we have in this environment and is most significant for the never-taker LATE upper bound.

8.3 The Choice of K

As a tuning parameter in the LP, we need to choose the order of Bernstein polynomials, K. In general, K should be chosen based on the sample size and the smoothness of the function to be approximated, in our case, $q(\cdot)$. The choice of the sieve dimension or more

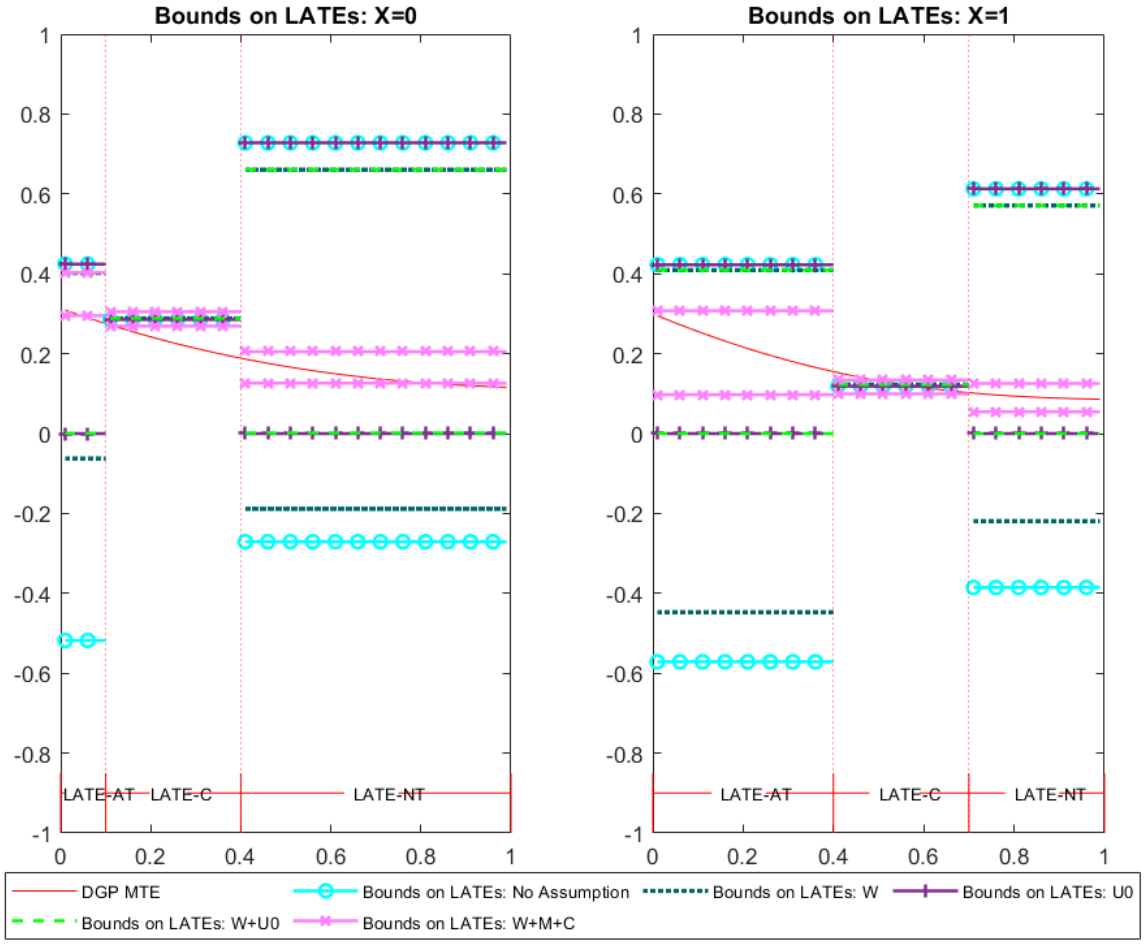


Figure 3: Bounds on the LATEs under Different Assumptions

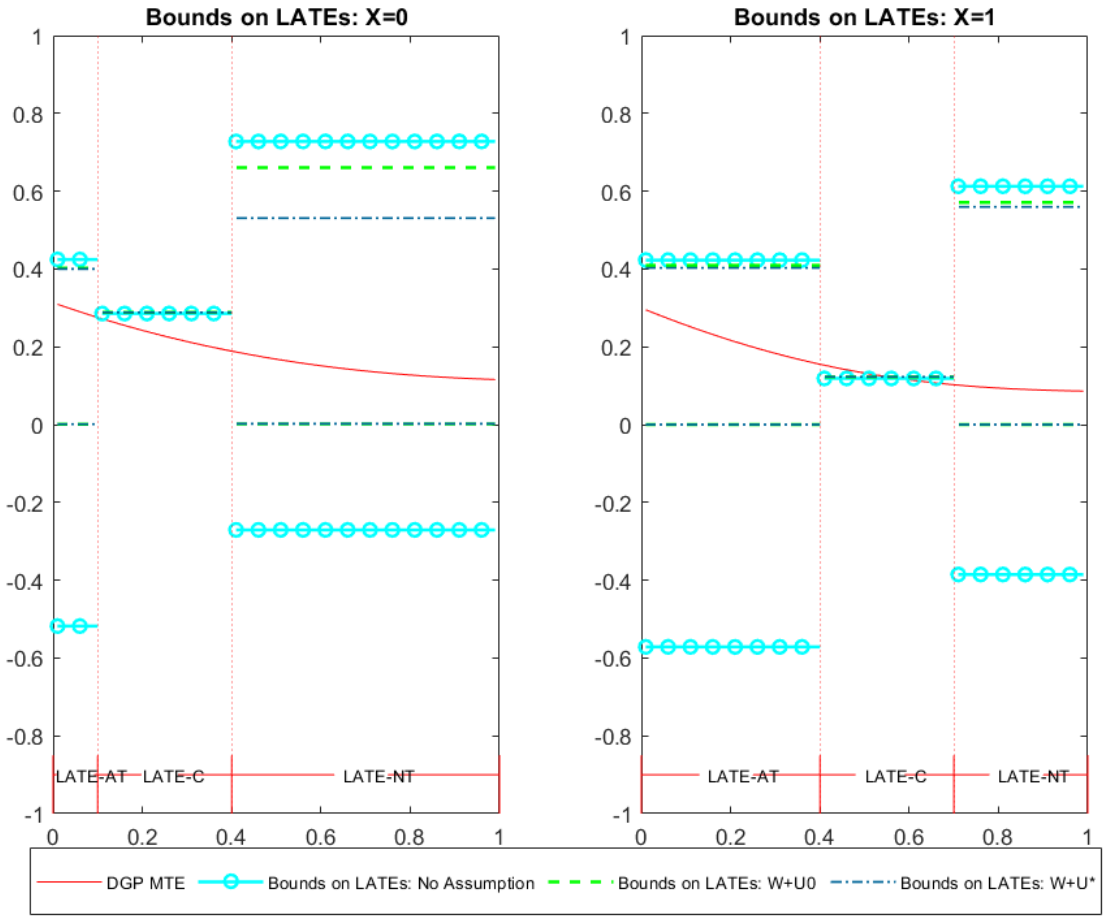


Figure 4: Bounds on the LATEs under Different Assumptions

generally, regularization parameters, is a difficult question (Chen (2007)) and developing data-driven procedure is a subject of on-going research in various nonparametric contexts of point identification; see, e.g., Chen and Christensen (2015) and Han (2020). In this partial identification setup, we propose the following heuristic and conservative approach, which is in spirit consistent with the very motivation of partial identification.

First, we do not want to claim any prior knowledge about the smoothness of $q(\cdot)$ because it is the distribution of a latent variable. Because K determines the dimension of unknown parameter θ in the linear programming, the width of the bounds tends to increase with K. At the same time, the computational burden increases with K. One interesting numerical finding is that, when K is sufficiently large, the increase of the width slows down and the bounds become stable. This suggests that we may be able to conservatively choose K that acknowledges our lack of knowledge of the smoothness but, at the same time, produces a reasonable computational task for the linear programming.

To illustrate this point, we consider the conditional MTE as the target parameter and show how its bounds change as K increases. We consider the MTE because it is a fundamental parameter that generates other target parameters, and hence, it is important to understand the sensitivity of its bounds to K. Figure 5 show the evolution of the bounds on the MTE as K grows. When $K = 5$, the bounds are narrow. Although it may be tempting to choose this value of K, this attempt should be avoided as it may be subject to the misspecification of smoothness. When K increases beyond 30, the bounds start to converge and become stable. We choose $K = 50$, and this is the choice we made in our previous numerical exercises.[9]

As discussed in Section B.2 in the Appendix, it is worth mentioning that the bounds on the MTE are point-wise sharp but *not* uniformly sharp. The graph for the MTE bounds are drawn by calculating the point-wise sharp bounds on MTE at each point of u (after properly discretizing it) and then connecting them. Therefore, these bounds should *not* be viewed as uniformly sharp bounds. Nonetheless, this graph is still useful for the purpose of our illustration. Given the current DGP, we find that there are no uniformly sharp bounds for the MTE.

[9]Note that with larger K, some LP solvers would ignore coefficients with negligible (e.g., 10^{-13}) values that cause a large range of magnitude in the coefficient matrix. It may be recommended to simultaneously rescale a column and a row to achieve a smaller range in the coefficients; see Section B.1 for details. When $K = 50$, the bounds from the rescaled LP and original LP are very close to each other (E.g., when we consider estimation of ATE without extra assumptions, the difference is up to 0.01). Therefore, setting K as large as 50 does not affect the main conclusion we have.

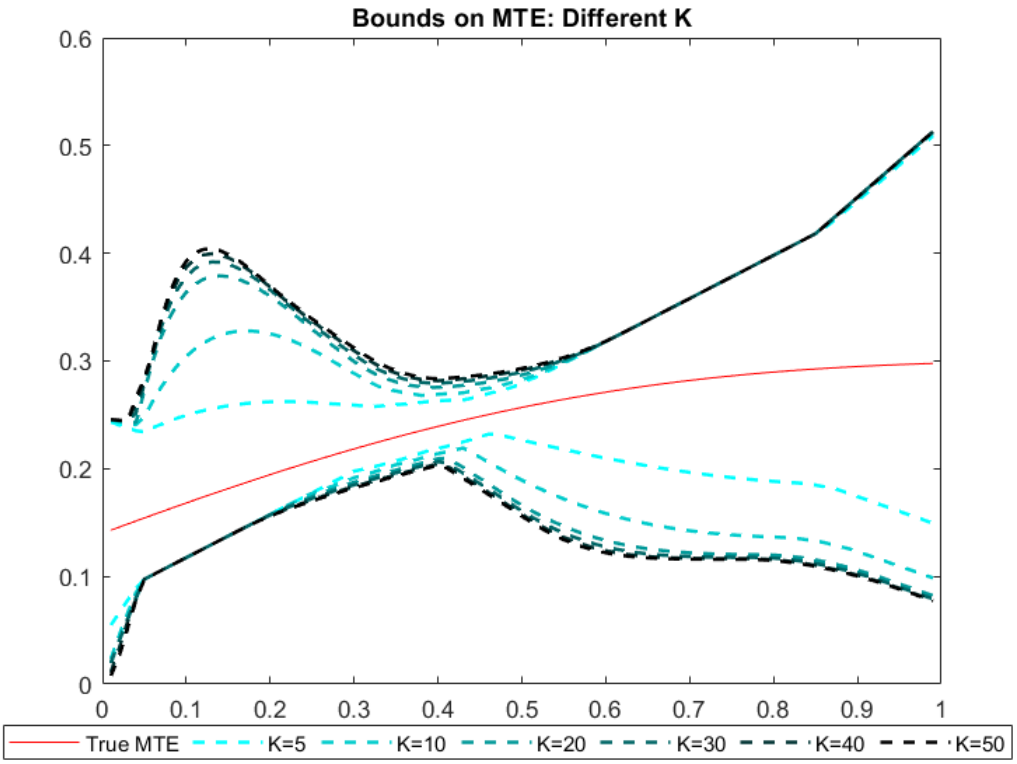


Figure 5: Bounds on MTE with Different K

9 Empirical Application

It is widely recognized in the empirical literature that health insurance coverage can be an essential factor for the utilization of medical services (Hurd and McGarry (1997); Dunlop et al. (2002); Finkelstein et al. (2012); Taubman et al. (2014)). Prior studies on this topic typically make use of parametric econometric models for the analysis. In their application, Han and Lee (2019) relax this common approach by introducing a semiparametric bivariate probit model to measure the average effect of insurance coverage on patients' medical visits. By applying our theoretical framework of partial identification, we further relax the parametric and semiparametric structures used in these studies. More importantly, we try to understand how much we can learn about the effect of insurance that is utilized through various counterfactual policies by learning the effect of different compliance groups.

We use the 2010 wave of the Medical Expenditure Panel Survey (MEPS) and focus on all the medical visits in January 2010. The sample is restricted to contain individuals aged between 25 and 64 and exclude those who had any kind of federal or state insurance in 2010. The outcome Y is a binary variable indicating whether or not an individual has visited a doctor's office; the treatment D is whether an individual has private insurance. We choose

whether a firm has multiple locations as the binary instrument Z. This IV reflects the size of the firm, and larger firms are more likely to provide fringe benefits, including health insurance. On the other hand, the number of branches of a firm does not directly affect employee decisions about medical visits. To justify the IV, self-employed individuals are excluded. For potentially endogenous covariates X, we include the age being 45 and older, gender, income above median, and health condition. Lastly, for an exogenous covariate W, we use the percentage of workers who are provided with paid sick leave benefits within each industry. Following Han and Lee (2019), we assume W satisfies Assumptions SEL_W(b) and EX_W(b), as X is controlled. We construct a categorical variable such that $W = 0$ for less than 50%, $W = 1$ for between 50–80%, and $W = 2$ for above 80%. Table 2 summarizes the observables.

Table 2: Summary Statistics

	Variable	Mean	S.D	Min	Max
Y	Whether or not visit doctors	0.18	0.39	0	1
D	Whether or not have insurance	0.66	0.47	0	1
Z	Firm has multiple locations	0.68	0.47	0	1
X	Age above 45	0.41	0.49	0	1
	Gender	0.50	0.50	0	1
	Income above median	0.50	0.50	0	1
	Good health	0.36	0.48	0	1
W	Pay sick leave provision	1.25	0.73	0	2
Number of observations = 7,555					

First, as a benchmark, we report that the LATE-C estimate calculated via our linear programming approach is equal to a singleton of 0.17, which is in fact identical to the 2SLS estimate we separately calculate. In what follows, we extrapolate this LATE beyond the complier group to the ATE. The presence of covariates reduces the effective sample size and thus leads to larger sampling errors in estimating the p of the ∞-LP (∞-LP1)–(∞-LP3). This may create inconsistencies in the set of equality constraints (∞-LP3), resulting in no feasible solution. This is in fact what happens in this application. To resolve this estimation problem, we introduce a slackness parameter η and modify (∞-LP3) so that, with some slackness, it satisfies

$$\|R_0 q - p\| \leq \eta. \tag{9.1}$$

A similarly modified constraint can then be followed in the finite-dimensional LP after ap-

proximation, as well as by combining (∞-LP4)–(∞-LP5). The appropriate value of η should depend on the sample size, the dimension of covariates, and the dimension of the unknown parameter θ. To explain the latter, as K increases, the dimension of θ (i.e., unknowns) increases, while the number of constraints (i.e., simultaneous equations for the unknowns) is fixed. Therefore, as K increases, the chance that the LP does not have a feasible solution would decrease. Based on the method discussed in the previous section, we set $K = 50$ in this application.

We calculate worst-case bounds on the ATE, as well as bounds after imposing Assumptions U^0 and M and after using covariate W. Under Assumption U^0, the data rules out the possibility that $Y(0) > Y(1)$, indicating that individuals with private insurance are more likely to visit a doctor. Assumption M imposes that the MTR function is weakly increasing in $U = u$. Usually, U is interpreted as the latent cost of obtaining treatment. Kowalski (2020) interpreted U as eligibility in a similar setup for Medicaid insurance. The eligibility for Medicaid is related to income level and age. In our setup, because the treatment is having the private insurance, we interpret the eligibility as the health status, which is reflected in the premium. Interpreting U as a latent cost (e.g., premium) of getting private insurance, Assumption M states that the chance of making a medical visit (with or without insurance) increases for those with higher cost. This is a reasonable assumption given that sicker individuals typically face higher insurance costs and also visit doctors more often. We choose the slackness parameter η to be 0.05 under no assumption and Assumption U and 0.07 when Assumption M is added. When W is used, we choose η to be 0.08 under no assumption and 0.1 with Assumption M.

The bounds on the ATE are shown in Figure 6. The worst-case bound on the ATE equals $[-0.45, 0.37]$. The bounds become $[0.01, 0.37]$ under Assumption U^0 and $[0.06, 0.37]$ under Assumption M. It is interesting to note that the identifying power of the uniformity and the shape restriction is similar in this example. When both Assumption U^0 and Assumption M are imposed, the bounds are further tightened to $[0.07, 0.37]$, although not substantially, indicating that the two assumptions are complementary. Lastly, we see improvements when the variation in W is exploited than when it is not, although the gains are not large.

Next, we consider the always-taker, complier, and never-taker LATEs. We consider these generalized LATEs conditional on $X = x$. Specifically, we focus on the treatment effects for males above age 45, with income below the median and bad health conditions. The results are shown in Table 3 and depicted in Figure 7. The LATE-C is analytically calculated via TSLS.[10] For the LATE-AT and LATE-NT, Assumption U^0 identifies the sign of the effects,

[10]When the alternative constraint (9.1) is used with the slackness parameter, the LATE-C is no longer a singleton.

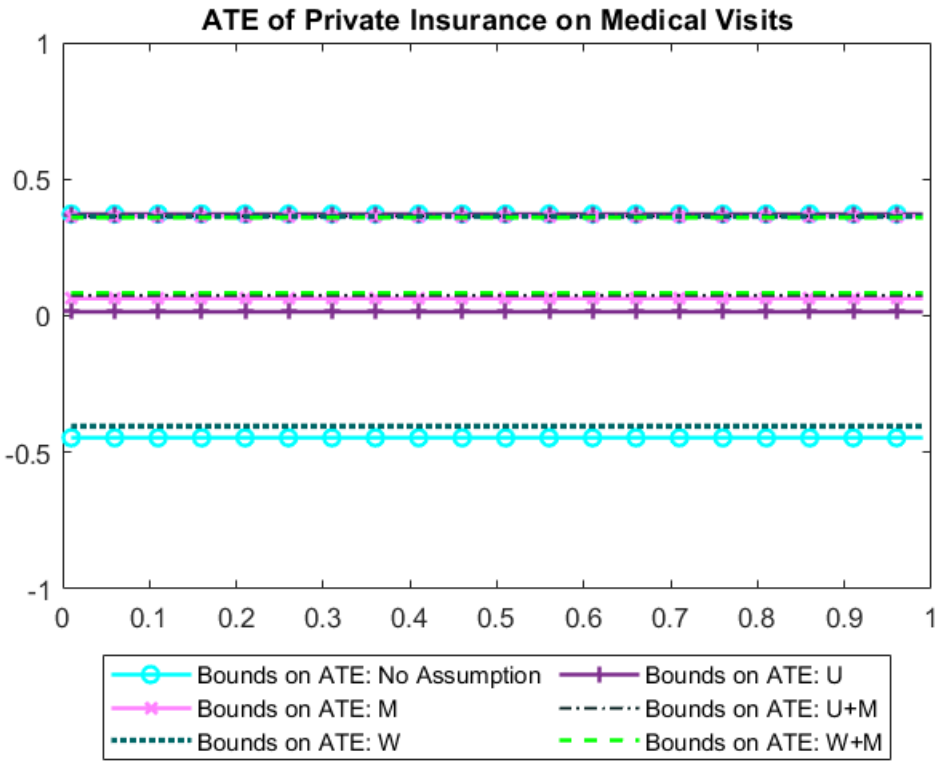


Figure 6: Bounds on the ATE of Private Insurance on Medical Visits

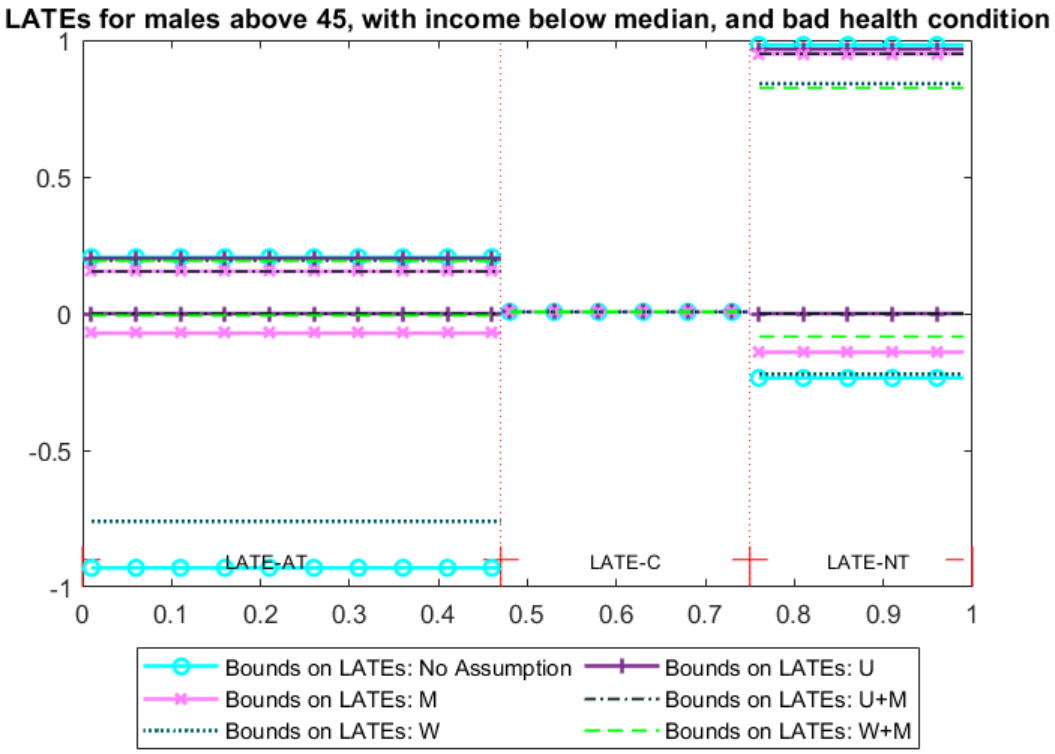


Figure 7: Bounds on the generalized LATEs of Private Insurance on Medical Visits for Male Above 45, with Income Below Median, of Bad Healthiness

Table 3: Estimated Bounds on generalized LATEs for Males Above 45, with Income Below Median, Bad Health Condition

	No Assumption	U^0	M	U^0+M	W	M+W
LATE-AT	[-0.93,0.21]	[0,0.20]	[-0.07,0.15]	[0,0.15]	[-0.76,0.20]	[-0.01,0.19]
LATE-C	0.01	0.01	0.01	0.01	0.01	0.01
LATE-NT	[-0.24,0.98]	[0,0.97]	[-0.14,0.95]	[0,0.95]	[-0.22,0.84]	[-0.08,0.82]
Slackness parameter η	0.05	0.05	0.07	0.07	0.08	0.10
Number of observations = 7,555						

and Assumption M nearly identifies it. Using the variation in W mostly improves the bounds compared to the ones without it.[11] From the results, we can conclude that possessing private insurance has the greatest effect on medical visits for never-takers, i.e., people who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because high costs might hinder those with the most need from receiving enough medical services.

A Examples of the Target Parameters

Table 4 contains the list of target parameters. The table is taken from Mogstad et al. (2018).

B More Discussions

B.1 Rescaling of Linear Programs

Let $B\theta = p$ represents the constraints (LP3) in the LP (LP1)–(LP3). In practice, the matrix B has the number of columns that grows with K. An important consequence is that, when K is large, the entries of B (i.e., constraint coefficients) take values of very different orders of magnitude; some coefficients are too small and some are too large. In this case, many optimization algorithms do not work properly and, to address the issue, some of them arbitrarily drop coefficients with small values (e.g., GUROBI drops coefficients that are less

[11] Most of the extra assumptions we impose help to determine the direction of treatment effect, i.e., to raise the lower bound if the treatment effect is positive. Therefore, improvements on LATE-NT are smaller than LATE-AT after imposing extra assumptions, since the evidence of positive treatment effect is relatively strong even with the worst-case bounds of LATE-NT.

Target Parameters	Expressions	Ranges of u	Weights
			$w_d(u,z,x)$
Average Treatment Effect (ATE)	$E[Y(1)-Y(0)]$	$[0,1]$	$2d-1$
LATE for Compliers (LATE-C) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [P(z_0,x),P(z_1,x)]\}$	$[P(z_0,x),P(z_1,x)]$	$(2d-1)\times \frac{1(u\in[P(z_0,x),P(z_1,x)])}{P(z_1,x)-P(z_0,x)}$
LATE for Always-Takers (LATE-AT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [0,P(z_0,x)]\}$	$[0,P(z_0,x)]$	$(2d-1)\times \frac{1(u\in[0,P(z_0,x)])}{P(z_0,x)}$
LATE for Never-Takers (LATE-NT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [P(z_1,x),1]\}$	$[P(z_1,x),1]$	$(2d-1)\times \frac{1(u\in[P(z_1,x),1])}{1-P(z_1,x)}$
LATE for $[\underline{u},\overline{u}]$	$E[Y(1)-Y(0)\|u \in [\underline{u},\overline{u}]]$	$[P(z_0,x),P(z_1,x)]$	$(2d-1)\times \frac{1(u\in[\underline{u},\overline{u}])}{\overline{u}-\underline{u}}$
Marginal Treatment Effect (MTE)*	$E[Y(1)-Y(0)\|u']$	u'	$(2d-1)\times 1(u=u')$
Policy Relevant Treatment Effect (PRTE) for a new policy (P',Z')	$\frac{E(Y')-E(Y)}{E(D')-E(D)}$	$[0,1]$	$(2d-1)\times \frac{\Pr[u\le P'(z')]-\Pr[u\le P'(z)]}{E[P(Z')]-E[P(Z)]}$

* The MTE uses the Dirac measure at u', while the other target parameters use the Lebesgue measure on $[0,1]$.

Table 4: Examples of the Target Parameters

than 10^{-13}). This may arbitrarily change the bounds we obtain. In this section, we propose a rescaling method to address this problem.

To better understand the rescaling strategy, we first express the original LP (LP1)–(LP3) in terms of matrices:

$$\max_{\theta\in\Theta_K} A\theta$$

subject to

$$B\theta = p.$$

Here θ is defined as a vector of unknown parameters $\{\theta_k^{e,x}\}_{k,e,x}$. And Θ_K is redefined as

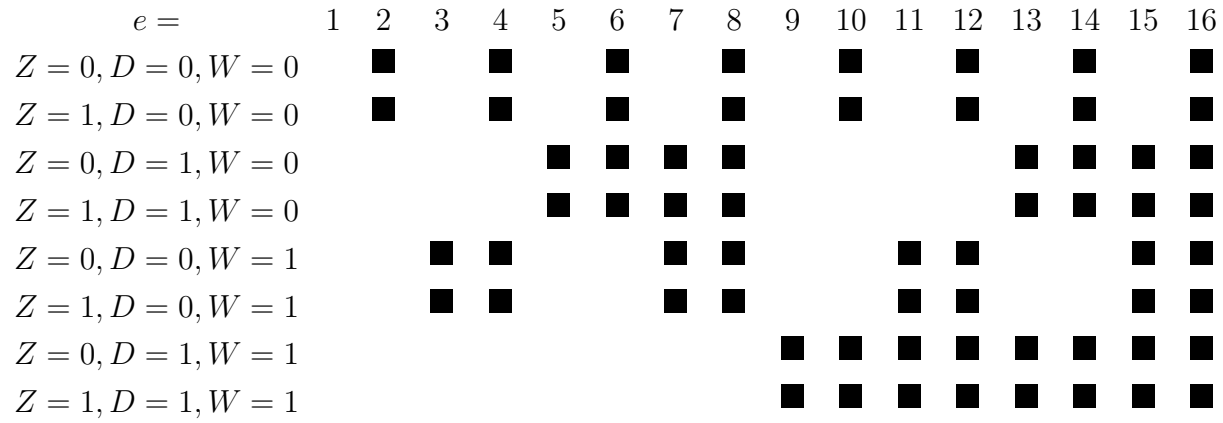
$$\Theta_K \equiv \{\theta : M\theta = \mathbf{1}, \theta \geq \mathbf{0}\},$$

where M is a weight matrix corresponding to $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1\ \forall(k,x)$, $\mathbf{1}$ is a column vector of ones, and $\mathbf{0}$ is a zero vector.

Because the Bernstein polynomials are only used in generating the coefficients in the equality data restrictions, we only consider rescaling of this constraint. Suppose the dimension of B is $m \times n$, and $m < n$.[12] We show that in our setting, B has full rank m. To show

[12] m is determined by the dimension of p, which is determined by the dimension of D, Y, Z, X, and n is determined by the order of polynomials we set in sieve approximation. Usually n is set to be a large number to guarantee the accuracy of sieve approximation, and it usually is larger than m. When $m = n$, theoretically we should achieve a unique solution, but empirically, the numerical error involved in the calculation process

this, we need to understand the structure of B. The number of columns of B is determined by the size of $\mathcal{E}$ and the order of polynomials K. The number of rows of B is determined by the dimension of p, i.e., the size of the support of (D, Y, Z, W, X). We consider an example with binary (D, Y, Z, W) for illustration. With binary (D, W), $|\mathcal{E}| = 16$ and B takes the form as below:



The square represents a vector of coefficients corresponding to θ's used to approximate the mapping types, and the blank represents a zero vector. By construction, each entry in matrix B is equivalent to $\int_{\mathcal{U}^d_{z,x}} b_{k,K}(u)du$ such that the multiplication of B and θ equals to the data distribution. From the matrix above, we can guarantee that B has full rank if, for fixed (d, w), the row representing $Z = 0$ cannot be a constant multiplication of the row representing $Z = 1$.

Lemma B.1. *Suppose $Z \in \{z_1, z_2\}$ is a binary instrument variable. Assume that $P(z_1), P(z_2) \in (0, 1)$ and $P(z_1) \neq P(z_2)$. For $k = 0, 1, \cdots K$, define $f(k) = \frac{\int_0^{P(z_1)} b_{k,K}(u)du}{\int_0^{P(z_2)} b_{k,K}(u)du}$. Then, $f(k)$ is not a constant function.*

Proof. Suppose $f(k)$ is a constant function, such that $f(k) = c$ for all k for some c. By properties of Bernstein polynomials (Farouki and Rajan (1988)), it satisfies that

$$\int_0^{P(z_1)} b_{k,K}(u)du = \frac{1}{K+1} \sum_{i=k+1}^{K+1} b_{i,K+1}\left(P(z_1)\right) = \frac{1}{K+1} \sum_{i=k+1}^{K+1} \left(\begin{array}{c} K+1 \\ i \end{array} \right) P(z_1)^i \left(1 - P(z_1)\right)^{K+1-i},$$

$$\int_0^{P(z_2)} b_{k,K}(u)du = \frac{1}{K+1} \sum_{i=k+1}^{K+1} b_{i,K+1}\left(P(z_2)\right) = \frac{1}{K+1} \sum_{i=k+1}^{K+1} \left(\begin{array}{c} K+1 \\ i \end{array} \right) P(z_2)^i \left(1 - P(z_2)\right)^{K+1-i}.$$

can result in infeasibility.

Let $k = m$ for some $m \in \{1, ..., K\}$. Then, $f(m) = c$ is equivalent to

$$\sum_{i=m+1}^{K+1} \frac{1}{K+1} \begin{pmatrix} K+1 \\ i \end{pmatrix} P(z_1)^i \left(1 - P(z_1)\right)^{K+1-i} = c \sum_{i=m+1}^{K+1} \begin{pmatrix} K+1 \\ i \end{pmatrix} \frac{1}{K+1} P(z_2)^i \left(1 - P(z_2)\right)^{K+1-i}. \tag{B.1}$$

Similarly let $k = m - 1$, then $f(m-1) = c$ is equivalent to

$$\sum_{i=m}^{K+1} \frac{1}{K+1} \begin{pmatrix} K+1 \\ i \end{pmatrix} P(z_1)^i \left(1 - P(z_1)\right)^{K+1-i} = c \sum_{i=m}^{K+1} \begin{pmatrix} K+1 \\ i \end{pmatrix} \frac{1}{K+1} P(z_2)^i \left(1 - P(z_2)\right)^{K+1-i}. \tag{B.2}$$

By subtracting (B.2) from (B.1), we have

$$\begin{pmatrix} K+1 \\ m \end{pmatrix} P(z_1)^m \left(1 - P(z_1)\right)^{K+1-m} = c \begin{pmatrix} K+1 \\ m \end{pmatrix} P(z_2)^m \left(1 - P(z_2)\right)^{K+1-m}$$

or equivalently,

$$c = \frac{P(z_1)^m \left(1 - P(z_1)\right)^{K+1-m}}{P(z_2)^m \left(1 - P(z_2)\right)^{K+1-m}}.$$

Because this equation holds for any $m \in \{1, ..., K\}$, take $m = 1$, then

$$c = \frac{P(z_1) \left(1 - P(z_1)\right)^K}{P(z_2) \left(1 - P(z_2)\right)^K} \tag{B.3}$$

and take $m = K$, then

$$c = \frac{P(z_1)^K \left(1 - P(z_1)\right)}{P(z_2)^K \left(1 - P(z_2)\right)}. \tag{B.4}$$

But (B.3) and (B.4) hold if and only if $P(z_1) = P(z_2)$, which is a contradiction. Analogously, we can show that $\tilde{f}(k) = \frac{\int_{P(z_1)}^1 b_{k,K}(u)du}{\int_{P(z_2)}^1 b_{k,K}(u)du}$ is not a constant function. Therefore, the coefficient matrix has full rank. □

Our goal is to rescale the coefficient matrix B to a new matrix $\tilde{B}$ such that its entries have orders of magnitude that are balanced. The most intuitive choice of $\tilde{B}$ is the fully reduced form of B. Note B usually has more rows than columns (otherwise, we achieve point identification), therefore, the fully reduced form of B would take the form of

$$\tilde{B} = [I, \mathbf{0}],$$

where I is the identity matrix of rank m, and $\mathbf{0}$ is a zero matrix with dimension $m \times (n-m)$. The next step is to find a transformation matrix X such that $BX = \tilde{B}$, so that we can rewrite the optimization problem as

$$\max_{\tilde{\theta} \in \tilde{\Theta}_K} \tilde{A}\tilde{\theta}$$

subject to

$$\tilde{B}\tilde{\theta} = p,$$

where $\tilde{\theta} = X^{-1}\theta$, $\tilde{A} = AX$, and $\tilde{\Theta}_K \equiv \left\{\tilde{\theta} : MX\tilde{\theta} = \mathbf{1}, X\tilde{\theta} \geq \mathbf{0}\right\}$.

We propose a simple algorithm to find a full rank X. Since B is the column-reduced form of B, X can be seen as the elementary operation matrix used to reach the reduced form. For simplicity, we use the transposed matrix B' and apply Gauss-Jordan elimination with partial pivoting on it to achieve row-reduced form, the exactly same procedure is also applied on an identity matrix I with dimension $n \times n$, and transpose what we get at the end to construct X. Because simple row operation reserves the rank, X is guaranteed to have full rank. Note that there may exist multiple solutions of X, which makes this problem computationally easier than solving an LP.

B.2 Point-wise and Uniform Sharp Bounds on MTE

In Section 2, we provided some examples of target parameters. The building block for these parameters is the MTE, $m_1(u) - m_0(u)$ (suppressing x). Heckman and Vytlacil (2005) show why this fundamental parameter can be of independent interest. Unlike other target parameters proposed here, we may want to allow the MTE to be a function of u (beyond evaluating it at a fixed u). In this section, we discuss the subtle issue of point-wise and uniform sharp bounds on $\tau_{MTE}(u) \equiv m_1(u) - m_0(u)$ as a function of u.

Suppress X for simplicity. Recall $q(u) \equiv \{q(e|u)\}_{e \in \mathcal{E}}$ and $\mathcal{Q} \equiv \{q(\cdot) : \sum_e q(e|u) = 1 \, \forall u \text{ and } q(e|u) \geq 0 \, \forall (e,u)\}$. Let $\mathcal{M}$ be the set of MTE functions, i.e.,

$$\mathcal{M} \equiv \left\{m_1(\cdot) - m_0(\cdot) : m_d(\cdot) = E[Y_d|U = \cdot] = \sum_{e \in \mathcal{E}: g_e(d)=1} q(e|\cdot) \, \forall d \in \{0,1\} \text{ for } q(\cdot) \in \mathcal{Q}\right\}.$$

The bounds on $\tau_{MTE} \in \mathcal{M}$ in the ∞-LP are given by using a Dirac delta function as a weight. Therefore, given evaluation point $u \in [0,1]$, (∞-LP1)–(∞-LP3) can be simplified as follows, defining the upper and lower bounds $\overline{\tau}(u)$ and $\underline{\tau}(u)$ (being explicit about the

evaluation point) on $\tau_{MTE}(u)$:

$$\overline{\tau}(u) = \sup_{q \in \mathcal{Q}} \sum_{e \in \mathcal{E}: g_e(1)=1} q(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q(e|u) \tag{B.5}$$

$$\underline{\tau}(u) = \inf_{q \in \mathcal{Q}} \sum_{e \in \mathcal{E}: g_e(1)=1} q(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q(e|u) \tag{B.6}$$

subject to

$$\sum_{e: g_e(d)=1} \int_{\mathcal{U}_z^d} q(e|\tilde{u}) d\tilde{u} = p(1, d|z) \qquad \forall (d, z) \in \{0, 1\}^2. \tag{B.7}$$

Then, for any fixed $u \in [0, 1]$,

$$\underline{\tau}(u) \leq \tau_{MTE}(u) \leq \overline{\tau}(u).$$

We argue that these bounds are point-wise sharp but not necessarily uniformly sharp for $\tau_{MTE}(\cdot)$.[13]

Definition B.1 (Point-wise Sharpness)**.** *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp if, for any $\bar{u} \in [0, 1]$, there exist $\overline{\tau}_{MTE,\bar{u}}, \underline{\tau}_{MTE,\bar{u}} \in \mathcal{M}$ such that $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ and $\underline{\tau}(\bar{u}) = \underline{\tau}_{MTE,\bar{u}}(\bar{u})$.*

Theorem B.1. *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp bounds on $\tau_{MTE}(\cdot)$.*

The proofs of this and other theorems appear later. Note that point-wise bounds will maintain some properties of an MTE function, but not all. For uniform sharpness, $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ themselves have to be MTE functions on $[0, 1]$, i.e., $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ should be elements in $\mathcal{M}$.

Definition B.2 (Uniform Sharpness)**.** *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are uniformly sharp if $\overline{\tau}(\cdot), \underline{\tau}(\cdot) \in \mathcal{M}$.*

The following theorem is almost immediate.

Theorem B.2. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $\overline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^*(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^*(e|u)$ for all $u \in [0, 1]$. Similarly, $\underline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $\underline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^\dagger(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^\dagger(e|u)$ for all $u \in [0, 1]$.*

The following is a more useful result that relates point-wise bounds with uniform bounds. For each $\bar{u}$, let $q^*_{\bar{u}}(\cdot)$ and $q^\dagger_{\bar{u}}(\cdot)$ be the point-wise maximizer and minimizer of (B.5)–(B.7), respectively.

[13] See Firpo and Ridder (2019) for related definitions of point-wise and uniform sharpness.

Corollary B.1. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $q^*_{\bar{u}}(\bar{u}) = q^*(\bar{u})$ for all $\bar{u} \in [0,1]$. Also, $\underline{\tau}(u)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $q^\dagger_{\bar{u}}(\bar{u}) = q^\dagger(\bar{u})$ for all $\bar{u} \in [0,1]$.*

Based on the Bernstein approximation we introduce, this corollary implies that for a uniform upper bound to exist, there should exist a common maximizer θ^* such that θ^* is in the feasible set of the LP and $\overline{\tau}(u) = \sum_{k \in \mathcal{K}} \left\{ \sum_{e \in \mathcal{E}: g_e(1)=1} \theta^{e*}_k b_k(u) - \sum_{e \in \mathcal{E}: g_e(0)=1} \theta^{e*}_k b_k(u) \right\}$ *for all u*. In other words, if $\theta^*_{\bar{u}}$ is the maximizer of the LP for given $\bar{u}$, then there should exist θ^* in the feasible set such that $\theta^*_{\bar{u}} = \theta^*$ for all $\bar{u} \in [0,1]$. Since this condition will not generally hold, uniformly sharp bounds on the MTE may not exist. The condition can be verified in practice by implementing the LP in a finite grid of u in $[0,1]$ and checking whether θ^*_u is constant for all values in the grid.

B.3 Linear Programming with Continuous X

Suppose X is continuously distributed and assume $\mathcal{X} = [0,1]^{d_X}$. Let $q(u,x) \equiv \{q(e|u,x)\}_{e \in \mathcal{E}}$ and $p(x) \equiv \{p(1,d|z,x)\}_{d,z}$. Recall that $R_\tau : \mathcal{Q} \to \mathbb{R}$ and $R : \mathcal{Q} \to \mathbb{R}^{d_p}$ are the linear operators of $q(\cdot)$ where d_p is the dimension of p. Consider the following LP:

$$\overline{\tau} = \sup_{q \in \mathcal{Q}} R_\tau q, \tag{B.8}$$

$$\underline{\tau} = \inf_{q \in \mathcal{Q}} R_\tau q, \tag{B.9}$$

$$s.t. \quad (Rq)(x) = p(x) \qquad \text{for all } x \in \mathcal{X}, \tag{B.10}$$

where $(Rq)(x) = p(x)$ emphasizes the dependence on x, and thus represents infinitely many constraints. Therefore, this LP is infinite dimensional because of both the decision variable and the constraints.

Now, for the sieve space of $\mathcal{Q}$, we consider

$$\tilde{\mathcal{Q}}_K \equiv \left\{ \left\{ \sum_{k=1}^{K} \theta^e_k b_k(u,x) \right\}_{e \in \mathcal{E}} : \sum_{e \in \mathcal{E}} \theta^e_k = 1 \text{ and } \theta^e_k \geq 0 \, \forall (e,k) \right\} \subseteq \mathcal{Q}, \tag{B.11}$$

where $b_k(u,x)$ is a bivariate Bernstein polynomial and $\mathcal{K} \equiv \{1,...,K\}$. Then,

$$\begin{aligned} E[\tau_d(Z,X)] &= \sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^e \int E[b_k(u,X) w_d(u,Z,X)]du \\ &\equiv \sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^e \tilde{\gamma}_k^d, \end{aligned} \tag{B.12}$$

where $\tilde{\gamma}_k^d \equiv \int E[b_k(u,X) w_d(u,Z,X)]du$. Also,

$$\begin{aligned} p(1,d|z,x) &= \sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^e \int_{\mathcal{U}_{z,x}^d} b_k(u,x)du \\ &\equiv \sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^e \tilde{\delta}_k^d(z,x), \end{aligned} \tag{B.13}$$

where $\tilde{\delta}_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u,x)du$. To deal with this infinite dimensional constraints (with respect to x), we proceed as follows. For any measurable function $h : \mathcal{X} \to \mathbb{R}$, $E\left|h(X)\right| = 0$ if and only if $h(x) = 0$ almost everywhere in $\mathcal{X}$. Therefore, the equality restriction (B.13) can be replaced by

$$E\left| \sum_{e:g_e(d)=y} \sum_{k\in\mathcal{K}} \theta_k^e \tilde{\delta}_k^d(z,X) - p(1,d|z,X) \right| = 0$$

for all $(d,z) \in \{0,1\}^2$. Let $\tilde{\theta} \equiv \{\theta_k^e\}_{(e,k)\in\mathcal{E}\times\mathcal{K}}$ and let

$$\tilde{\Theta}_K \equiv \left\{ \tilde{\theta} : \sum_{e\in\mathcal{E}} \theta_k^e = 1 \text{ and } \theta_k^e \geq 0 \, \forall (e,k) \in \mathcal{E}\times\mathcal{K} \right\}.$$

Then, we can formulate the following finite-dimensional LP:

$$\overline{\tau}_K = \max_{\theta\in\Theta_K} \sum_{k\in\mathcal{K}} \left\{ \sum_{e:g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e:g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \right\} \tag{B.14}$$

$$\underline{\tau}_K = \min_{\theta\in\Theta_K} \sum_{k\in\mathcal{K}} \left\{ \sum_{e:g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e:g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \right\} \tag{B.15}$$

subject to

$$E\left| \sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^e \tilde{\delta}_k^d(z,X) - p(1,d|z,X) \right| = 0 \qquad \forall (d,z) \in \{0,1\}^2. \tag{B.16}$$

Later, we want to introduce additional constraints from some identifying assumptions:

$$R_1 q = a_1 \tag{B.17}$$

$$R_2 q \leq a_2 \tag{B.18}$$

For the equality restrictions, we can use the same approach that transforms (B.10). For the inequality restrictions (B.18), we can allow any identifying assumptions for which R_2 is a matrix rather than an operator:

Assumption MAT. *R_2 is a $\dim(a_2) \times \dim(q)$ matrix.*

Assumptions M and C and the unconditional version of Assumption MTS satisfy this condition.

B.4 Estimation and Inference

Although the paper's main focus is identification, we briefly discuss estimation and inference. The estimation of the bounds characterized by the LP (LP1)–(LP3) is straightforward by replacing the population objects $(\gamma_k^d, \delta_k^d, p)$ with their sample counterparts $(\hat{\gamma}_k^d, \hat{\delta}_k^d, \hat{p})$. With continuous Y in (7.6)–(7.8), we replace (7.8) with its sample counterparts and a slack version:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\sum_{e:g_e(d)=1}\sum_{k\in\mathcal{K}}\theta_k^e\hat{\tilde{\delta}}_k^d(Z_i,X_i) - \hat{p}(1,d|Z_i,X_i)\right| \leq \eta,$$

where $\hat{p}(1,d|z,x)$ is some preliminary estimate of $p(1,d|z,x)$ and η is the slackness parameter. A similar idea applies to the case with continuous X in (B.14)–(B.16).

It is important to construct a confidence set for our target parameter or its bounds in order to account for the sampling variation in measuring treatment effectiveness. It will also be interesting to develop a procedure to conduct a specification test for the identifying assumptions discussed in Section 6. The problem of statistical inference when the identified set is constructed via linear programming has been studied in, e.g., Deb et al. (2017), Mogstad et al. (2018), Hsieh et al. (2018), Torgovitsky (2019b), and Fang et al. (2020). Among these papers, Mogstad et al. (2017)'s setting is closest to our setting with discrete variables, and their inference procedure can be directly adapted to our problem. Instead of repeating their result here, we only briefly discuss the procedure.

Recall $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ is the latent distribution and $p \equiv \{p(1,d|z,x)\}_{d,z,x}$ is the distribution of the data, and R_τ, R_0, R_1, and R_2 denote the linear operators of $q(\cdot)$ that

correspond to the target and constraints. Consider the following hypotheses:

$$H_0 : p \in \mathcal{P}_0, \qquad H_1 : p \in \mathcal{P} \backslash \mathcal{P}_0,$$

where

$$\mathcal{P}_0 \equiv \{p \in \mathcal{P} : Rq = a \text{ for some } q \in \mathcal{Q}\}$$

and

$$R \equiv (R_\tau', R_0', R_1', R_2')'$$
$$a \equiv (\tau, p', a_1', a_2')'$$

Suppose $\hat{R}$ and $\hat{a}$ are sample counterparts of R and a. Then, a minimum distance test statistic can be constructed as

$$T_n(\tau) \equiv \inf_{q \in \mathcal{Q}_K} \sqrt{n} \left\| \hat{R}q - \hat{a} \right\|.$$

Similar to Mogstad et al. (2017), $T_n(\tau)$ is the solution to a convex optimization problem that can be reformulated as an LP using duality. A $(1-\alpha)$-confidence set for the target parameter τ can be constructed by inverting the test:

$$CS_{1-\alpha} \equiv \{\tau : T_n(\tau) \leq \hat{c}_{1-\alpha}\}$$

where $\hat{c}_{1-\alpha}$ is the critical value for the test. The resulting object is of independent interest, and it can further be used to conduct specification tests. The large sample theory for $T_n(\tau)$, as well as a bootstrap procedure to calculate $\hat{c}_{1-\alpha}$, will directly follow according to Mogstad et al. (2017), which is omitted for succinctness.

When Y or X is continuously distributed, then the resulting LP is semi-infinite dimensional. In this case, the inference procedure by Chernozhukov et al. (2013) can be applied. In this case, the estimation of the bounds can be conducted within the framework.

B.5 Equivalence with the IV-Like Estimands

We draw a connection between our approach and the approach used in Mogstad et al. (2018) in the case of binary Y. In particular, we show that the identified set of the MTR functions $\mathcal{M}_{id}$ used in Mogstad et al. (2018) is equivalent to the set of MTR functions derived from the feasible set used in this paper. Therefore, the feasible set in this paper contains no less

information about the data than those contained in $\mathcal{M}_{id}$ via IV-like estimands in their paper.

The IV-like estimand is defined in Proposition 3 in Mogstad et al. (2018), and is stated as below.

Proposition B.1 (IV-like Estimand from Mogstad et al. (2018))**.** *Suppose that* $s : \{0,1\} \times \mathbf{R}^{d_z \times d_x} \to \mathbf{R}$ *is an identified (or known) function that is measurable and has a finite second moment. We refer to such a function* s *as an IV-like specification and to* $\beta_s \equiv E\Big[s(D,Z,X)Y\Big]$ *as an IV-like estimand. If* (Y,D) *are generated according to Assumption SEL and Assumption EX, then*

$$\beta_s = E\Big[\int_0^1 m_0(u,X)\omega_{0s}(u,Z,X)du\Big] + E\Big[\int_0^1 m_1(u,X)\omega_{1s}(u,Z,X)du\Big], \tag{B.19}$$

where $\omega_{0s}(u,z,x) = s(0,z,x)1[u > p(z,x)]$*, and* $\omega_{1s}(u,z,x) = s(1,z,x)1[u \leq p(z,x)]$*.*

For the MTR functions to be consistent with the data, the following conditions need to be satisfied:

$$E[Y|D=0,Z,X] = E[Y_0|U > p(Z,X),Z,X] = \frac{1}{1-P(Z,X)}\int_{p(Z,X)}^1 m_0(u,X)du, \tag{B.20}$$

$$E[Y|D=1,Z,X] = E[Y_1|U \leq p(Z,X),Z,X] = \frac{1}{P(Z,X)}\int_0^{p(Z,X)} m_1(u,X)du. \tag{B.21}$$

Define the identified set as:

$$\mathcal{M}_{id} = \Big\{m = (m_0,m_1), m_0, m_1 \in L^2 : m_0, m_1 \text{ satisfies equation (B.20) and (B.21) a.s}\Big\}.$$

This identified set is defined in Mogstad et al. (2018, Section 2.5). The definition follows the fact that the MTR functions in $\mathcal{M}_{id}$ are compatible with the observed conditional means of Y. In this sense, it exhausts the information of the data contained in the conditional means. When Y is binary, the conditional means of Y contain the information of the complete distribution.

Define the feasible set $\mathcal{Q}_f$ as

$$\mathcal{Q}_f = \Big\{q \in L^2 : q \in \mathcal{Q} \text{ and satisfies equation } (\infty\text{-LP3})\Big\}.$$

To establish the connection with $\mathcal{M}_{id}$, we construct the set of MTR functions based on the feasible set:

$$\mathcal{M}_f = \Big\{ m = (m_0, m_1) : m_d = \sum_{e: g_e(d)=1} q(e|u,x), d = \{0,1\}, q \in \mathcal{Q}_f \Big\}.$$

Then the following holds, proof of which appears later:

Theorem B.3. *Suppose Y is discretely distributed. Under the Assumption SEL and EX, $\mathcal{M}_f = \mathcal{M}_{id}$.*

Proposition 3 in Mogstad et al. (2018) shows an equivalence relationship between the identified set $\mathcal{M}_{id}$ and the set of MTR functions satisfying constraints based on selected IV-like estimands. Theorem B.3 shows that the information contained in our feasible set used in the LP is the same as the selected IV-like estimands that exhaust the available information. Theorem B.3 can be extended to the case where Y is discrete and X is continuous. When Y is a non-binary discrete or continuous outcome, $\mathcal{M}_{id}$ and $\mathcal{M}_f$ only exhaust the information on the conditional means, but not other distributional information. Nonetheless, that missing information is captured by $\mathcal{Q}_f$ that we use as our constraint set, because $q(e|u)$ is defined as the conditional probability of Y taking each value.

C Proofs

C.1 Proof of Lemma 4.1

Fix (d, z, x). By $\sum_{e \in \mathcal{E}} q(e|u,x) = 1$ for $q \in \mathcal{Q}$, we have

$$1 = \sum_{e \in \mathcal{E}} q(e|u,x) = \sum_{e: g_e(d)=1} q(e|u,x) + \sum_{e: g_e(d)=0} q(e|u,x).$$

Then, in (∞-LP3), the constraint with $p(0,d|z,x)$ can be written as

$$\begin{aligned} p(0,d|z,x) &= \int_{\mathcal{U}^d_{z,x}} \sum_{e: g_e(d)=0} q(e|u,x) du = \int_{\mathcal{U}^d_{z,x}} \Big\{ 1 - \sum_{e: g_e(d)=1} q(e|u,x) \Big\} du \\ &= \Pr[D = d | Z = z, X = x] - \int_{\mathcal{U}^d_{z,x}} \sum_{e: g_e(d)=1} q(e|u,x) du. \end{aligned}$$

Then by rearranging terms, this constraint becomes

$$p(1,d|z,x) = \int_{\mathcal{U}^d_{z,x}} \sum_{e: g_e(d)=1} q(e|u,x) du,$$

since $\Pr[D = d|Z = z, X = x] - p(0, d|z, x) = p(1, d|z, x)$. Therefore, the constraint with $p(0, d|z, x)$ does not contribute to the restrictions imposed by (∞-LP3) and $q \in \mathcal{Q}$. □

C.2 Proof of Theorem 6.1

In proving the claim of the theorem for W, we fix $Z = z$. We first prove with Case (a). To simplify notation, let $q(e_1, ..., e_J|u) \equiv \Pr[\epsilon \in \{e_1, ..., e_J\}|u] = \sum_{j=1}^{J} q(e_j|u)$. Based on Table (1), we can easily derive

$$
\begin{aligned}
p(1,1|z,1) &= \int_0^{P(z)} \sum_{e:g_e(1,1)=1} q(e|u)du = \int_0^{P(z)} q(9,...,16|u)du, \\
p(1,1|z,0) &= \int_0^{P(z)} \sum_{e:g_e(1,0)=1} q(e|u)du = \int_0^{P(z)} q(5,...,8,13,...,16|u)du, \\
p(1,0|z,1) &= \int_{P(z)}^1 \sum_{e:g_e(0,1)=1} q(e|u)du = \int_{P(z)}^1 q(3,4,7,8,11,12,15,16|u)du, \\
p(1,0|z,0) &= \int_{P(z)}^1 \sum_{e:g_e(0,0)=1} q(e|u)du = \int_{P(z)}^1 q(2,4,6,8,10,12,14,16|u)du.
\end{aligned}
$$

Define the operator

$$
T_z^d q^e \equiv \int_{\mathcal{U}_z^d} q(e|u)du.
$$

Then, for the r.h.s. $(p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'$ of the constraints in (LP3) that correspond to $Z = z$, the corresponding l.h.s. is

$$
\begin{aligned}
&\begin{pmatrix}
\int_0^{P(z)} q(9,...,16|u)du \\
\int_0^{P(z)} q(5,...,8,13,...,16|u)du \\
\int_{P(z)}^1 q(3,4,7,8,11,12,15,16|u)du \\
\int_{P(z)}^1 q(2,4,6,8,10,12,14,16|u)du
\end{pmatrix} \\
&= \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\
0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\
0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 \\
0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0
\end{pmatrix} q \\
&\equiv Tq,
\end{aligned}
$$

where T is a matrix of operators implicitly defined and $q(u) \equiv (q(1|u), ..., q(16|u))$. Now for $q \in \mathcal{Q}_K$, define a $16K$-vector

$$\theta \equiv \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^{16} \end{pmatrix}$$

where, for each $e \in \{1, ..., 16\}$, $\theta^e \equiv (\theta^e_1, ..., \theta^e_K)'$. Similarly, let $b(u) \equiv (b_1(u), ..., b_K(u))'$. Then, we have $q(e|u) = b(u)'\theta^e$. Let H be a 16×16 diagonal matrix of 1's and 0's that imposes additional identifying assumptions on the outcome data-generating process. In this proof, H is used to incorporate Assumption R(i). Given H, the constraints in (LP3) (that correspond to $Z = z$) can be written as

$$THq = \{TH \otimes b'\}\theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Now, we prove the claim of the theorem. Suppose the claim is not true, i.e., the even rows are linearly dependent to odd rows in TH. Given the form of T, which has full rank under Assumption R(ii)(a), this linear dependence only occurs when H is such that $H_{jj} = 1$ for $j \in \{1, 4, 13, 16\}$ and 0 otherwise. But, according to Table 1, this implies that $\Pr[Y(d, w) \neq Y(d, w')] = 0$ for all d and $w \neq w'$, which contradicts Assumption R(i). This proves the theorem for Case (a).

Now we move to prove the theorem for Case (b), analogous to the previous case. For every z, we can derive

$$\begin{aligned}
p(1, 1|z, 1) &= \int_0^{P(z,1)} \sum_{e:g_e(1,1)=1} q(e|u)du = \int_0^{P(z,1)} q(9, ..., 16|u)du, \\
p(1, 1|z, 0) &= \int_0^{P(z,0)} \sum_{e:g_e(1,0)=1} q(e|u)du = \int_0^{P(z,0)} q(5, ..., 8, 13, ..., 16|u)du, \\
p(1, 0|z, 1) &= \int_{P(z,1)}^1 \sum_{e:g_e(0,1)=1} q(e|u)du = \int_{P(z,1)}^1 q(3, 4, 7, 8, 11, 12, 15, 16|u)du, \\
p(1, 0|z, 0) &= \int_{P(z,0)}^1 \sum_{e:g_e(0,0)=1} q(e|u)du = \int_{P(z,0)}^1 q(2, 4, 6, 8, 10, 12, 14, 16|u)du.
\end{aligned}$$

Define

$$T^d_{z,w} q^e \equiv \int_{\mathcal{U}^d_{z,w}} q(e|u)du$$

where $\mathcal{U}^d_{z,w}$ can be analogously defined. Then,

$$\begin{pmatrix} \int_0^{P(z,w)} q(9,...,16|u)du \\ \int_0^{P(z,w')} q(5,...,8,13,...,16|u)du \\ \int_{P(z,w)}^1 q(3,4,7,8,11,12,15,16|u)du \\ \int_{P(z,w')}^1 q(2,4,6,8,10,12,14,16|u)du \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} \\ 0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & 0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} \\ 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} \\ 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} \end{pmatrix} q$$
$$\equiv \tilde{T}q,$$

where $\tilde{T}$ is a matrix of operators implicitly defined. Then, inserting H, the constraint becomes

$$\tilde{T}Hq = \left\{\tilde{T}H \otimes b'\right\}\theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Then the remaining argument is the same as in the previous case, which completes the proof for W. The proof for Z can be analogously done and is more straightforward, so is omitted. □

C.3 Proof of Theorem B.1

For any given $\bar{u} \in [0,1]$, $\overline{\tau}(\bar{u}) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$ for some $q^*_{\bar{u}}(\cdot) \equiv \{q^*_{\bar{u}}(e|\cdot)\}_{e\in\mathcal{E}}$ in the feasible set of the LP, (B.5) and (B.7). Therefore, $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ for $\overline{\tau}_{MTE,\bar{u}}(\bar{u}) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$, which is in $\mathcal{M}$ by definition. We can have a symmetric proof for $\underline{\tau}(\cdot)$. □

C.4 Proof of Theorem B.2

Again, by the fact that $\tau_{MTE}(\cdot) = \sum_{e\in\mathcal{E}:g_e(1)=1} q(e|\cdot) - \sum_{e\in\mathcal{E}:g_e(0)=1} q(e|\cdot)$ in general, $\overline{\tau}(u) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*(e|u) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*(e|u)$ for all $u \in [0,1]$ is equivalent to $\overline{\tau}(\cdot)$ being contained in $\mathcal{M}$, and similarly for $\underline{\tau}(\cdot)$. □

C.5 Proof of Theorem B.3

From (∞-LP3), we can write $E[Y|D=0,Z,X]$ in terms of $q(e|u,X)$ as below:

$$
\begin{aligned}
E[Y|D=0,Z,X] &= \Pr\left[Y=1|D=0,Z,X\right] = \frac{\Pr\left[Y=1,D=0|Z,X\right]}{\Pr\left[D=0|Z,X\right]} \\
&= \frac{1}{1-P(Z,X)} \sum_{e:g_e(0)=1} \int_{P(Z,X)}^{1} q(e|u,X)du \\
&= \frac{1}{1-P(Z,X)} \int_{P(Z,X)}^{1} \sum_{e:g_e(0)=1} q(e|u,X)du \qquad \text{(C.1)}
\end{aligned}
$$

Therefore, for $(m_0, m_1) \in \mathcal{M}_f$

$$
E[Y|D=0,Z,X] = \frac{1}{P(Z,X)} \int_{P(Z,X)}^{1} m_0(u,X)du
$$

and symmetrically,

$$
E[Y|D=1,Z,X] = \frac{1}{P(Z,X)} \int_{0}^{P(Z,X)} m_1(u,X)du
$$

We conclude that $\mathcal{M}_f \subset \mathcal{M}_{id}$.

Now suppose $m \in \mathcal{M}_{id}$. By (B.20) and (C.1), for $\forall z, x$

$$
\frac{1}{1-P(z,x)} \int_{P(z,x)}^{1} m_0(u,x)du = \frac{1}{1-P(z,x)} \sum_{e:g_e(0)=1} \int_{P(z,x)}^{1} q(e|u,x)du
$$

and,

$$
\int_{P(z,x)}^{1} \left[m_0(u,x) - \sum_{e:g_e(0)=1} q(e|u,x) \right] du = 0
$$

This equality holds for all the possible values of $P(z,x)$, we conclude that $m_0(u,x) = \sum_{e:g_e(0)=1} q(e|u,x)$ on the support $u \in [0,1]$, $\forall x$ following the fundamental theorem of calculus. Following the symmetric procedure, we can conclude that $m_1(u,x) = \sum_{e:g_e(1)=1} q(e|u,x)$. And we show that $\mathcal{M}_{id} \subset \mathcal{M}_f$. Thus, $\mathcal{M}_f = \mathcal{M}_{id}$.

References

Angrist, J. and I. Fernandez-Val (2010): "Extrapolate-ing: External validity and overidentification in the late framework," Tech. rep., National Bureau of Economic Research. 1

Balat, J. F. and S. Han (2018): "Multiple treatments with strategic interaction," *Available at SSRN 3182766*. 1, 2

Balke, A. and J. Pearl (1997): "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 1, 4

Bertanha, M. and G. W. Imbens (2019): "External validity in fuzzy regression discontinuity designs," *Journal of Business & Economic Statistics*, 1–39. 1

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008): "Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization," *American Economic Review*, 98, 351–56. 6

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125, 985–1039. 1, 2, 3.3

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632. 8.3

Chen, X. and T. Christensen (2015): "Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation," . 8.3

Chen, X., E. T. Tamer, and A. Torgovitsky (2011): "Sensitivity analysis in semiparametric likelihood models," . 5

Chen, X., J. Tan, Z. Liu, and J. Xie (2017): "Approximation of functions by a new family of generalized Bernstein operators," *Journal of Mathematical Analysis and Applications*, 450, 244–261. 5

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261. 1, 3.1

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81, 667–737. B.4

Coolidge, J. L. (1949): "The story of the binomial theorem," *The American Mathematical Monthly*, 56, 147–157. 5

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: Alternative methods for the evaluation of policy interventions," *Labour Economics*, 41, 47–60. 2

Deb, R., Y. Kitamura, J. K.-H. Quah, and J. Stoye (2017): "Revealed price preference: Theory and stochastic testing," . B.4

Dehejia, R., C. Pop-Eleches, and C. Samii (2019): "From local to global: External validity in a fertility natural experiment," *Journal of Business & Economic Statistics*, 1–27. 1

Dunlop, D. D., L. M. Manheim, J. Song, and R. W. Chang (2002): "Gender and ethnic/racial disparities in health care utilization among older adults," *The Journals of Gerontology Series B: Psychological sciences and social sciences*, 57, S221–S233. 9

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The generalized Roy model and the cost-benefit analysis of social programs," *Journal of Political Economy*, 123, 413–443. 2

Fang, Z., A. Santos, A. Shaikh, and A. Torgovitsky (2020): "Inference for large-scale linear systems with known coefficients," *University of Chicago, Becker Friedman Institute for Economics Working Paper*. B.4

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012): "The Oregon health insurance experiment: evidence from the first year," *The Quarterly journal of economics*, 127, 1057–1106. 9

Firpo, S. and G. Ridder (2019): "Partial identification of the treatment effect distribution and its functionals," *Journal of Econometrics*, 213, 210–234. 13

Gunsilius, F. (2019): "Bounds in continuous instrumental variable models," *arXiv preprint arXiv:1910.09502*. 1

Han, S. (2019): "Optimal dynamic treatment regimes and partial welfare ordering," *arXiv preprint arXiv:1912.10014*. 1, 6

——— (2020): "Nonparametric estimation of triangular simultaneous equations models under weak identification," *Quantitative Economics*, 11, 161–202. 8.3

——— (2021): "Identification in nonparametric models for dynamic treatment effects," *Journal of Econometrics*, 225, 132–147. 5

Han, S. and S. Lee (2019): "Estimation in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Applied Econometrics*, 34, 994–1015. 1, 2, 9

Han, S. and E. J. Vytlacil (2017): "Identification in a generalization of bivariate probit models with dummy endogenous regressors," *Journal of Econometrics*, 199, 63–73. 1, 2

Heckman, J. J. (2010): "Building bridges between structural and program evaluation approaches to evaluating policy," *Journal of Economic literature*, 48, 356–98. 1

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation1," *Econometrica*, 73, 669–738. 1, 2, 2, B.2

Heckman, J. J. and E. J. Vytlacil (1999): "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the national Academy of Sciences*, 96, 4730–4734. 1

Hsieh, Y.-W., X. Shi, and M. Shum (2018): "Inference on estimators defined by mathematical programming," *Available at SSRN 3041040*. B.4

Hurd, M. D. and K. McGarry (1997): "Medical insurance and the use of health care services by the elderly," *Journal of Health Economics*, 16, 129–154. 9

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 2

Joy, K. I. (2000): "Bernstein polynomials," *On-Line Geometric Modeling Notes*, 13. 5

Kamat, V. (2019): "Identification with latent choice sets: The case of the head start impact study," *arXiv preprint arXiv:1711.02048*. 1

Kowalski, A. E. (2020): "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," Tech. rep., National Bureau of Economic Research. 1, 2, 9

Machado, C., A. Shaikh, and E. Vytlacil (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212, 522–555. 1

Manski, C. F. (1997): "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1311–1334. 3.1

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. 1, 3.1, 3.2

Marx, P. (2020): "Sharp Bounds in the Latent Index Selection Model," *arXiv preprint arXiv:2012.02390*. 1, 3.1, 5

Masten, M. A. and A. Poirier (2018): "Salvaging falsified instrumental variable models," *arXiv preprint arXiv:1812.11598*. 5

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects," Tech. rep., National Bureau of Economic Research. B.4

——— (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. 1, 2, 4, 2, 3.3, 4, 5, 5, 6, 6, 8, A, B.4, B.5, B.1, B.5, B.5

Mogstad, M., A. Torgovitsky, and C. R. Walters (2019): "Identification of causal effects with multiple instruments: Problems and some solutions," Tech. rep., National Bureau of Economic Research. 5

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81. 1, 2

Muralidharan, K., A. Singh, and A. J. Ganimian (2019): "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 109, 1426–60. 1

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial identification in triangular systems of equations with binary dependent variables," *Econometrica*, 79, 949–955. 1, 2, 6

Taubman, S. L., H. L. Allen, B. J. Wright, K. Baicker, and A. N. Finkelstein (2014): "Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment," *Science*, 343, 263–268. 9

Torgovitsky, A. (2019a): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505. 1

——— (2019b): "Nonparametric inference on state dependence in unemployment," *Econometrica*, 87, 1475–1505. B.4

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. 1, 2

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70, 331–341. 2

Vytlacil, E. and N. Yildiz (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, 75, 757–779. 1, 2