Partial Identification in Nonseparable Binary Response Models with Endogenous Regressors

Jiaying Gu[*] *University of Toronto*

Thomas M. Russell[†] *Carleton University*

July 19, 2021

Abstract

This paper considers (partial) identification of a variety of counterfactual parameters in binary response models with possibly endogenous regressors. Our framework allows for nonseparable index functions with multi-dimensional latent variables, and does not require parametric distributional assumptions. We leverage results on hyperplane arrangements and cell enumeration from the literature on computational geometry in order to provide a tractable means of computing the identified set. We demonstrate how various functional form, independence, and monotonicity assumptions can be imposed as constraints in our optimization procedure to tighten the identified set, and we show how these assumptions can be assigned meaningful interpretations in terms of restrictions on latent response types. Finally, we apply our method to study the effects of health insurance on the decision to seek medical treatment.

Keywords: Binary Choice, Counterfactual Probabilities, Endogeneity, Hyperplane Arrangement, Linear Programming, Partial Identification

We are grateful to Marc Henry, Roger Koenker, and the seminar audiences at Columbia University and Michigan State University for helpful feedback. We also thank Martin Weidner and the organizers of the Chamberlain Seminar, and are grateful to Florian Gunsilius, Sukjin Han, Wayne Gao, and Takuya Ura for their questions and feedback, and to Adam Rosen for his thoughtful discussion. Jiaying Gu acknowledges financial support from the Social Sciences and Humanities Research Council of Canada. All errors are our own.

[*] Jiaying Gu, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S3G7, Canada. Email: jiaying.gu@utoronto.ca.

[†] Thomas M. Russell, Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S5B6, Canada. Email: thomas.russell3@carleton.ca.

1 Introduction

This paper considers partial identification of a variety of parameters in a general class of binary response models. Our main focus throughout is on counterfactual probabilities, as well as parameters that can be written as linear combinations of counterfactual probabilities. Our approach focuses on the case when the observed random variables have finite support and the index function is a linear function of the latent variables, but otherwise we allow for flexible functional form assumptions, endogenous regressors, and the inclusion of multi-dimensional and nonseparable latent variables. Furthermore, our approach does not require any parametric distributional assumptions.

In the settings closest to the one we consider, nonparametric point-identification of the distribution of latent variables occurs only under restrictive conditions, often including strong independence assumptions and large support conditions (e.g. Ichimura and Thompson (1998)). Control function approaches are often used to address the issue of endogenous regressors, but if endogenous regressors are discrete or the mechanism generating the endogenous regressors is poorly understood, then many of these approaches are not applicable. Partial identification arises as a natural alternative to methods for point-identification as a result of possible endogeneity, discrete instruments, and limited variation in the covariates. However, flexible and easily implementable methods in partial identification for binary response models remain underdeveloped. This paper seeks to address this gap.

Our analysis reveals the importance of a special partition of the latent variable space into *types* that have identical responses in all possible counterfactual states. Consistent with the previous literature, we call these latent types *response types*. We show that additional functional form, independence, and monotonicity assumptions can all be assigned meaning in terms of these response types. In particular, we show that functional form and monotonicity assumptions are equivalent to the *elimination of response types*, which amounts to assigning zero probability to regions of the latent variable space corresponding to a particular profile of counterfactual responses. Furthermore, we show that certain independence assumptions imposed on a vector of latent variables are observationally equivalent to imposing independence on response types directly. This connection helps to facilitate interpretation of these assumptions in the class of models we consider. We are not the first to emphasize the importance of response types, and our discussion echoes the insights of Balke and Pearl (1994), Heckman and Pinto (2018), and many others.

One of our main contributions is to show how to enumerate the set of all response types consistent with a given binary choice model using the cell enumeration algorithm of Gu and Koenker (2020) for hyperplane arrangements, a required step before bounding counterfactual quantities. In particular, we show that linearity of the index function in the latent variables naturally partitions the latent variable space into cells defined by a collection of hyperplanes, where each cell corresponds to a unique response type. Using the cell enumeration algorithm of Gu and Koenker (2020), we show how to enumerate all response types that are consistent with a given collection of assumptions in polynomial time. After enumerating all response types, we show how various counterfactual quantities can be easily bounded by solving two linear programming problems. Our procedure thus represents a feasible method of constructing bounds on counterfactual quantities under weak assumptions where the latent variables may be multi-dimensional and nonseparable.

We also thoroughly study the special case when the index function is linear in parameters. In contrast to most existing procedures in partial identification, we show that exact (i.e. not approximate) sharp bounds on counterfactual probabilities can be computed without the need to grid over the entire parameter space. Our procedure leverages results in convex analysis, and combines the double description algorithm of Fukuda and Prodon (1995) with the cell enumeration algorithm of Gu and Koenker (2020). Avoiding the need to grid over the entire parameter space allows us to derive a plug-in estimator consistency result that is valid under weak assumptions, and also allows us to easily adapt the recent inference procedure of Cho and Russell (2020) to the setting in this paper to efficiently construct confidence sets and bias-corrected estimates of the identified set.

Finally we apply our method to study the effects of private health insurance on the decision to seek medical treatment. Consistent with the existing literature, we treat private health insurance status as an endogenous variable, and we consider the decision to seek medical treatment as our binary outcome variable of interest. We then consider the average treatment effect of obtaining private health insurance on the decision to visit a doctor. We find that the sign of the average treatment effect is typically only identified under our strongest assumptions. However, even our strongest assumptions are much weaker than the assumptions typically maintained in the empirical literature. Interestingly, we find non-trivial bounds on the average treatment effect even when the structural parameters are unidentified.

1.1 Review of Relevant Literature

Binary response models with possibly endogenous regressors have been studied extensively, and previous work on the subject can be separated into two broad categories: work that focuses on conditions required for point identification, and work that allows for partial identification. From the point identification perspective, typical approaches include (i) the use of linear probability models, (ii) maximum likelihood estimation (e.g. the bivariate probit), and (iii) control function approaches. All of these approaches have well-documented limitations.[1] In particular, linear probability models are commonly justified as approximations to the underlying conditional expectation function for the binary dependent variable, but are known to deliver misleading results when the conditional expectation function is highly nonlinear.[2] Methods that use maximum likelihood—such as the bivariate probit model—enjoy efficiency gains relative to other approaches when the model is correctly specified, but require strong *a priori* knowledge of the mechanism generating the endogenous variables, as well as knowledge of the joint distribution of the latent variables up to some finite parameter vector. Finally, control function approaches (e.g. Blundell and Smith (1989), Blundell and

[1] A review of approaches typically used by practitioners to address the problem of endogenous regressors in models with binary outcomes is provided in Lewbel et al. (2012), who focus on the case of a threshold-crossing model with linear index function and additively separable errors.

[2] Lewbel et al. (2012) construct an interesting treatment effect example with a binary outcome variable where the treatment effect is positive for everyone, but the ATE under a linear probability model is negative.

Powell (2004), and Imbens and Newey (2009), among many others) relax (to some extent) the assumptions required on the latent variables, but are generally restricted to cases with continuous endogenous variables and still require a correctly specified model for the endogenous variables in nonlinear models. Unlike the control function approach, the special regressor approach of Lewbel (2000) (see also Lewbel et al. (2012) and Dong and Lewbel (2015)) does not require the correct specification of a model for endogenous variables, but instead requires the existence of an observed continuously distributed regressor with large support that satisfies certain conditional independence assumptions. Such a special regressor is not always readily available.

Beyond these approaches, a number of papers have considered nonparametric identification. Nonparametric identification was studied in binary choice and threshold crossing models by Matzkin (1992), and in more general nonseparable models by Matzkin (2003) and Chernozhukov and Hansen (2005), among others. Vytlacil and Yildiz (2007) study nonparametric identification of the average treatment effect in a discrete triangular system with a binary endogenous variable under a weak separability assumption in the outcome equation. Important precedents to the work presented here from the literature on point identification in random coefficient models include Ichimura and Thompson (1998), Gautier and Kitamura (2013) and Gu and Koenker (2020). However, all these papers focus almost exclusively on the point-identified case with linear index function and exogenous covariates with large support.

In contrast, the literature on partial identification attempts to relax the assumptions required for point-identification. In a relevant series of papers, Chesher et al. (2013), Chesher and Rosen (2014) show how to use random set theory to characterize the identified set of structures in discrete choice models. A general formulation of their approach is presented in Chesher and Rosen (2017). Similar to the current paper, these papers do not provide a model for the endogenous covariates, rendering the discrete choice model *incomplete*. Chesher et al. (2013) and Chesher and Rosen (2014) then use a characterization of the sharp set of constraints given by a result due to Artstein (1983) in random set theory.[3] Our work extends the work by Chesher et al. (2013) and Chesher and Rosen (2014), although we focus on obtaining sharp bounds on a general class of counterfactual conditional distributions, and show how this can be accomplished by solving a sequence of optimization problems with the help of results from computational geometry.

In other relevant precedents to our work, Galichon and Henry (2011), Laffers (2019) and Torgovitsky (2019) demonstrate how to construct sharp bounds on various parameters in models with finite variables by appropriately partitioning the latent variable space and discretizing the latent variables. We also use an identification argument based on partitioning the latent variable space. However, the current paper focuses substantially on how to practically compute the relevant partition using results from the literature on computational geometry, which is otherwise nontrivial with more than one or two latent variables. We also show how to avoid griding over the entire parameter space when computing the identified set. This allows us to compute *exact* (i.e. not approximate) bounds in our model, and also allows us to easily modify a

[3] See also Norberg (1992) and Molchanov (2017) Corollary 1.4.11.

recent and simple subvector/functional inference procedure (Cho and Russell (2020)) to efficiently compute confidence sets and biased-corrected estimates of the identified set.

There are a number of other relevant papers in the literature on partial identification in discrete choice models. In an important paper, Manski (2007) considers counterfactual choice probabilities in a setting with partial identification, and shows how these counterfactual choice probabilities can be bounded using optimization problems. However, the general approach used in this paper is very different. Furthermore, we focus substantially on demonstrating how to practically incorporate a flexible set of assumptions on the latent index function, and we allow for endogenous explanatory variables.[4] In another related and recent working paper, Tebaldi et al. (2019) study the problem of computing various counterfactual quantities in a nonparametric discrete choice model with an application to consumer choice of health insurance in California. However, they focus specifically on the case where consumers have quasi-linear utility functions (equal to their valuation of the insurance option minus the premium) and use the particular structure of their setting and problem to resolve issues of endogeneity by conditioning on a set of covariates.[5] Computational considerations—a major contribution of the current paper—are also not addressed in these papers.

Closely related to the problem of bounding counterfactual probabilities is the problem of bounding parameters in the literature on treatment effects with binary outcome variables. Analytic bounds in triangular systems of equations with binary dependent variables under various assumptions is considered by Chiburis (2010), Shaikh and Vytlacil (2011), and Mourifié (2015). An optimization-based approach to bounding treatment effect parameters is presented in Russell (2019) in the discrete case, and Gunsilius (2020) in the continuous case.

This paper also makes a connection to the literature on computational geometry. In the case of a linear index function, computation of our bounds requires the analysis of a partition of the latent space determined by finitely many hyperplanes. This turns out to be a well studied subject in combinatorial geometry, and leads us to consideration of the enumeration algorithm proposed by Gu and Koenker (2020).

1.2 Paper Outline and Notation

The remainder of the paper proceeds as follows. Section 2 introduces the main theoretical framework and main assumptions. Section 3 studies practical implementation of the theoretical framework from Section 2 and introduces our optimization-based bounding procedure for counterfactual probabilities. Section 4 then demonstrates how to introduce functional form, independence, and monotonicity assumptions into our bounding procedure, and Section 5 applies our methodology to study the impact of health insurance on utilization of health care services. Section 6 concludes. All proofs can be found in Appendix A. Appendix B

[4] This latter point differentiates our work from Chiong et al. (2017) and Allen and Rehbeck (2019).

[5] In particular, Tebaldi et al. (2019) consider a multinomial choice model with preferences over insurance options given by the difference between the consumer's latent valuation and the consumer's premium for each option. Endogeneity arises because of possible dependence between valuations and premiums. However, in their setting (subsidized) premiums are deterministic functions of the coverage area, age, and income. The authors then discretize age and income, and assume that a valuation distribution is fixed within a given coverage area and discretized age and income bin; the remaining variation in premiums within each coverage area and discretized age and income bin is then considered to be exogenous.

provides some additional discussion of the results presented in the main text.

Notation: The following notation is relevant for both the main text and the appendices. Given a subset $\mathcal{X}$ of Euclidean space, we use $\mathfrak{B}(\mathcal{X})$ to denote the Borel $\sigma-$algebra on $\mathcal{X}$. For two measurable spaces $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ and $(\mathcal{X}', \mathfrak{B}(\mathcal{X}'))$, the product $\sigma-$algebra on $\mathcal{X} \times \mathcal{X}'$ is denoted by $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{X}')$. Random variables are denoted using capital letters, and if $X : (\Omega, \mathfrak{A}) \to (\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is a random variable defined on the probability space $(\Omega, \mathfrak{A}, P)$, then we use P_X to denote the probability measure induced on $\mathcal{X}$ by X; that is, for any $A \in \mathfrak{B}(\mathcal{X})$, $P_X(A) := P(X^{-1}(A))$. Furthermore, we interpret $P_{X|X'}(X \in A \mid X' = x')$ as a regular conditional probability measure. Finally, $P_{X|X'}$ is used as shorthand for the collection $P_{X|X'} := \{P_{X|X'}(\,\cdot \mid X' = x') : x' \in \mathcal{X}'\}$. We do not explicitly differentiate between scalars and vectors, or random variables and random vectors. To keep the notation clean, we sometimes omit the transpose when combining column vectors; that is, if v_1 and v_2 are two column vectors, rather than write $v = (v_1^\top, v_2^\top)^\top$ we instead write $v = (v_1, v_2)$, where it is understood that v is a column vector unless otherwise specified.

2 General Framework: Theoretical Considerations

In this section we first introduce our main assumptions on the binary response model, and connect our assumptions to the definition of the identified set of (conditional) latent variable distributions. We then turn to the problem of bounding counterfactual parameters.

2.1 The Identified Set of Latent Variable Distributions

We start by introducing our main assumptions on the binary response environment under consideration.

Assumption 2.1. *There exists a complete probability space $(\Omega, \mathfrak{A}, P)$, a random variable $Y : \Omega \to \{0, 1\}$, and random vectors $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}^{d_x}$, $Z : \Omega \to \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ and $U : \Omega \to \mathcal{U} = \mathbb{R}^{d_u}$ satisfying:*

$$Y = \mathbb{1}\{\varphi(X, Z, U, \theta_0) \geq 0\} \text{ a.s.}, \tag{2.1}$$

for some function $\varphi(\,\cdot\,, \theta_0) : \mathcal{X} \times \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$ parameterized by $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ with:

$$\varphi(\,\cdot\,, \theta) = \tilde{\varphi}_1(x, z, \theta)^\top u + \tilde{\varphi}_2(x, z, \theta), \tag{2.2}$$

where $\tilde{\varphi}_1(\,\cdot\,, \theta)$ and $\tilde{\varphi}_2(\,\cdot\,, \theta)$ are measurable for each θ. Furthermore, $|\mathcal{X}| = m_x < \infty$, and $|\mathcal{Z}| = m_z < \infty$, the spaces $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{U}$ are equipped with the Borel $\sigma-$algebra, and the distribution of U is absolutely continuous with respect to the Lebesgue measure.

In Assumption 2.1 $U \in \mathcal{U}$ is a vector of latent variables, $\theta \in \Theta$ is a vector of fixed coefficients, and $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ and $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ are vectors of covariates. From (2.2) we restrict the index function to be linear in the latent variables $U \in \mathcal{U}$, although the model in Assumption 2.1 still allows for general nonseparability between covariates and latent variables. In this model the latent variables can also be interpreted as random coefficients, in which case there is no restriction on which covariates are assigned fixed versus random coefficients by the index function φ. A special case of linearity occurs when the function φ is additively separable in a scalar latent variable U, which occurs, for instance, when $\tilde{\varphi}_1(x, z, \theta) = 1$. A full analysis of this special case using the framework in this paper is taken up in Appendix B.5. For now there is no distinction between X and Z, and either may be dependent with the latent vector U. Throughout the paper we switch freely between indexing the points in $\mathcal{X} \times \mathcal{Z}$ either by $\{(x_1, z_1), (x_1, z_2), \ldots, (x_{m_x}, z_{m_z})\}$ or by $\{(x_1, z_1), (x_2, z_2), \ldots, (x_m, z_m)\}$ with $m := m_x \cdot m_z$, depending on which method is more convenient for our purpose. Finally, imposing absolute continuity of the distribution of latent variables is standard in this literature, and although it is not required for any of the major results it allows for a dramatic simplification of the cell enumeration algorithm introduced in the next section.

We assume that the researcher's objective throughout is to obtain a sharp set of constraints defining the identified set of latent variable distributions, and to use these constraints to bound various counterfactual quantities, such as counterfactual conditional probabilities. Similar to previous works, we take the *selection* relation as a primitive relation on which to construct a definition of the identified set. The close connection between the selection relation from random set theory and the concept of *observational equivalence* from the work in econometrics on identification has been appreciated in Beresteanu et al. (2011), Beresteanu et al. (2012), Chesher et al. (2013), Chesher and Rosen (2014), and Chesher and Rosen (2017), among many others. We continue this work here. In particular, we define the set:

$$\mathcal{U}(y, x, z, \theta) := \{u \in \mathcal{U} : y = \mathbb{1}\{\varphi(x, z, u, \theta) \geq 0\}\}. \tag{2.3}$$

Chesher and Rosen (2017) call this set the $U-$level set. Intuitively, (2.3) delivers all possible values of the latent variables u consistent with the vector (y, x, z, θ) given the binary response model in (2.1). A *measurable selection* from the random set $\mathcal{U}(Y, X, Z, \theta)$ is a random vector $U : \Omega \to \mathcal{U}$ satisfying $U \in \mathcal{U}(Y, X, Z, \theta)$ a.s.[6] Importantly, given a distribution of the observable random vectors (Y, X, Z), a structural function φ and a fixed coefficient $\theta \in \Theta$, any two measurable selections U and U' from the random set $\mathcal{U}(Y, X, Z, \theta)$ are *observationally equivalent* in the sense that both latent variable vectors U and U' are consistent with the observed distribution of Y, X and Z for the vector of parameters $\theta \in \Theta$ through the model (2.1). Framed in this manner, constructing the identified set of latent variable distributions then becomes a problem of verifying whether a given random vector $U : \Omega \to \mathcal{U}$ is a measurable selection from the random set in (2.3), and then collecting the distributions of all such selections.

We now present the definition of the joint identified set for the (conditional) latent variable distribution and coefficients θ.

[6] A general definition of a selection and a random set is provided in Appendix A.2. In Appendix A.2 we prove that $\mathcal{U}(Y, X, Z, \theta)$ is suitably measurable and thus is a random set under our assumptions (see Lemma A.1). We also prove that $\mathcal{U}(Y, X, Z, \theta)$ admits a universally measurable selection (see Lemma A.2).

Definition 2.1 (Identified Set)**.** *Under Assumption 2.1, the (joint) identified set $\mathcal{I}^*_{Y,X,Z}$ of conditional latent variable distributions $P_{U|Y,X,Z}$ and fixed coefficients θ is the set of all pairs $(P_{U|Y,X,Z}, \theta)$ satisfying:*

$$P_{U|Y,X,Z}(U \in \mathcal{U}(Y,X,Z,\theta) \mid Y = y, X = x, Z = z) = 1, \;\; P_{Y,X,Z} - a.s. \tag{2.4}$$

Note that this definition of the identified set implicitly depends on the distribution of (Y,X,Z) through the almost-sure relation in (2.4); any values of (y,x,z) assigned zero probability by the observed distribution do not impose any restrictions on the distribution of U. Importantly, the definition conditions on the value of the endogenous outcome variable Y. This conditioning is carried throughout the paper, and we show in Section 5 that it allows us to bound some interesting, albeit less-typical counterfactual parameters that may be relevant to policy analysis. Definition 2.1 can also be used to define other related identified sets, including identified sets for conditional latent variable distributions of the form $P_{U|X,Z}$, $P_{U|Z}$, or P_U.

2.2 Bounding Counterfactual Quantities

In this paper, we limit ourselves to a class of counterfactual queries that can be characterized by the occurrence of an *intervention*. An intervention is represented by an exogenous causal process capable of manipulating the values of X and Z. For exogenous random variables—that is, those whose values are determined outside of the model—we simply replace the random variable by its value under consideration in the counterfactual. For endogenous random variables—that is, those whose values are determined by a function of the other exogenous and endogenous variables within a model—the function determining the value of the endogenous variable is deleted from the system, and the endogenous variable is replaced by its value under consideration in the counterfactual.[7] The following assumption summarizes this discussion.

Assumption 2.2 (Counterfactual Domain)**.** *For some collection of functions Γ with typical element $\gamma : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$, there exists a collection of random variables $\{Y(\,\cdot\,,\gamma) : \Omega \to \{0,1\} \mid \gamma \in \Gamma\}$, abbreviated as $Y_\gamma := Y(\,\cdot\,,\gamma)$, representing counterfactual choices for each γ such that $Y_\gamma : \Omega \to \{0,1\}$ is measurable for each γ, and:*

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta_0) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = 1,$$

$P_{Y,X,Z,U}-a.s.$ for the same $\theta_0 \in \Theta$ as in Assumption 2.1, and for all $\gamma \in \Gamma$.

Assumption 2.2 implies that (i) counterfactual response variables indexed by $\gamma \in \Gamma$ exist on the common probability space from Assumption 2.1, and (ii) such counterfactual response variables are equal (almost surely) to the values that would arise after an intervention on the system represented by (2.1). Under Assumption 2.2 each counterfactual is represented by a function $\gamma : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$ belonging to the collection Γ. Taking γ as a function allows us to consider a general class of counterfactuals that allows the

[7] We refer the reader to Pearl (2009) Section 7.1 for a discussion of a similar procedure. Such counterfactuals have a natural interpretation as "hypothetical experiments," and are widely attributed to Haavelmo (1943, 1944).

counterfactual under consideration to depend on the observed values of X and Z. Although each function γ is seen as a map from $\mathcal{X} \times \mathcal{Z}$ to itself, this does not prevent consideration of counterfactuals where γ selects values of (x, z) that have never been observed in the data. Such cases can be accommodated by simply extending the support $\mathcal{X} \times \mathcal{Z}$ from Assumption 2.1 to include any counterfactual pair (x, z) of interest.[8]

Assumption 2.2 on the counterfactual domain leads directly to our definition of the identified set for counterfactual conditional distributions.

Definition 2.2 (Identified Set of Counterfactual Conditional Distributions)**.** *Under Assumptions 2.1 and 2.2, the identified set of counterfactual conditional distributions $\mathcal{P}^*_{Y_\gamma|Y,X,Z,U}$ is the set of all conditional distributions $P_{Y_\gamma|Y,X,Z,U}$ satisfying:*

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = 1,$$

$P_{Y,X,Z,U}-a.s.$ for some $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$.

Note that this definition makes an explicit reference to the identified set $\mathcal{I}^*_{Y,X,Z}$ presented in Definition 2.1, which in turn is derived from a selection relation. As was the case with Definition 2.1, this definition of the identified set can be used as a starting point to define other related identified sets, including for counterfactual distributions of the form $P_{Y_\gamma|Y,X,Z}$ or $P_{Y_\gamma|X,Z}$, as well as identified sets for average structural functions and average treatment effects.

Using Definition 2.1, the following result provides an intuitive but important link between counterfactual distributions and the conditional distribution of latent variables.

Theorem 2.1. *Suppose that Assumptions 2.1 and 2.2 hold. Then a counterfactual conditional distribution $P_{Y_\gamma|Y,X,Z}$ satisfies $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$ if and only if there exists a pair $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying:*

$$P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z),U,\theta) \geq 0 \mid Y = y, X = x, Z = z\right), \tag{2.5}$$

$P_{Y,X,Z}-a.s.$

Theorem 2.1 provides the theoretical link between the identified set of counterfactual conditional distributions, and the identified set for the pair $(P_{U|Y,X,Z}, \theta)$. While the result is theoretically straightforward, it hides some important practical difficulties that arise when constructing the identified set for counterfactual conditional distributions. In particular, verifying the existence of a pair $(P_{U|Y,X,Z}, \theta)$ that satisfies the conditions from Definition 2.1 is a nontrivial task. This is at least partly due to the fact that $P_{U|Y,X,Z}$ is an infinite dimensional object, even in the case when both X and Z have finite support. This *infinite dimensional existence problem* is exacerbated in practice by the fact that $P_{U|Y,X,Z}$ must satisfy a number of constraints to ensure it is consistent with the binary response model through (2.4), and to ensure it is a proper conditional probability measure. We consider these practical difficulties in detail in the next section.

[8] This approach does not affect anything we present in this paper, since we always require any relation to the observed distribution of (Y, X, Z) to hold only almost-surely.

3 General Framework: Practical Considerations

In order to bound counterfactual probabilities using Theorem 2.1, we must verify the existence of a collection of Borel probability measures on $\mathcal{U}$ that are consistent with the binary response model through (2.4). However, solving this existence problem by explicitly constructing a probability measure on all Borel sets of $\mathcal{U}$ seems excessively difficult and naive. Instead, we would like to consider a finite collection of Borel sets that are both necessary and sufficient for this existence problem in the sense that, to solve the existence problem, it is both necessary and sufficient that we are able to construct a conditional probability measure on our finite collection of sets.[9]

To make progress, let us define the following vector-valued function:

$$r(u,\theta) := \begin{bmatrix} \mathbb{1}\{\varphi(x_1,z_1,u,\theta)\geq 0\} & \mathbb{1}\{\varphi(x_1,z_2,u,\theta)\geq 0\} & \dots & \mathbb{1}\{\varphi(x_{m_x},z_{m_z},u,\theta)\geq 0\}\end{bmatrix}^\top, \tag{3.1}$$

and for a fixed binary vector $s \in \{0,1\}^m$ let us define the set:

$$\mathcal{U}(s,\theta) := \{u \in \mathcal{U} : r(u,\theta) = s\}\,. \tag{3.2}$$

The sets from (3.2) partition the space $\mathcal{U}$ into at most 2^m sets, with each set being uniquely associated with a binary vector $s \in \{0,1\}^m$.[10] Similar objects to $r(u,\theta)$ have appeared previously in the literature (e.g. Balke and Pearl (1994), Heckman and Pinto (2018)), and to remain consistent with the previous literature we call the functions $r : \mathcal{U}\times\Theta \to \{0,1\}^m$ defined in (3.1) *response types.*[11] In the discrete choice setting, these response types tell us the choices that an individual with type indexed by (u,θ) *would have made* had they been assigned alternate pairs of (x,z). Any two individuals characterized by values of u from the same set $\mathcal{U}(s,\theta)$ make identical choices in every counterfactual, and so the values of u define a natural equivalence class of latent types.

After partitioning the space of latent variables using response types, various counterfactual objects of interest can be written as a disjoint union of the sets $\mathcal{U}(s,\theta)$ from (3.2) that comprise our partition. For the sake of illustration, consider the binary vectors:

$$S_j = \{s \in \{0,1\}^m : s_j = 1\}, \tag{3.3}$$

for $j = 1,\dots,m$. Note that each set S_j is comprised of all binary vectors that have a j^{th} entry equal to 1,

[9] A similar problem is addressed in Torgovitsky (2019), although we note that his general framework is not immediately applicable here since we are dealing with probability measures rather than distribution functions. We find that for many of the models we consider, it is simply not possible to write the identified set and functional of interest in terms of the multi-dimensional distribution function for the latent variables.

[10] This comes from the fact that there are m points of support in $\mathcal{X}\times\mathcal{Z}$ (and so m rows in $r(u,\theta)$) and each row of $r(u,\theta)$ can take values either 0 or 1.

[11] The collection of sets defining response types appears to be similar to the "minimal relevant partition" in Tebaldi et al. (2019), as well as the partition described in Chesher and Rosen (2014) Appendix B.

and thus contain exactly 2^{m-1} elements.[12] Now note, by definition of the sets $\mathcal{U}(s,\theta)$ and S_j we have:

$$\{u \in \mathcal{U} : \varphi(x_j, z_j, u, \theta) \geq 0\} = \bigcup_{s \in S_j} \mathcal{U}(s,\theta).$$

Furthermore, for $s' \neq s$ the definition of the sets $\mathcal{U}(s,\theta)$ from (3.2) ensures we have $\mathcal{U}(\theta, s') \cap \mathcal{U}(s,\theta) = \varnothing$, so that the union in the previous display is a disjoint union. Thus, we have the following decomposition:

$$P_{U|Y,X,Z}\left(\varphi(x_j, z_j, u, \theta) \geq 0 \mid Y = y, X = x, Z = z\right) = \sum_{s \in S_j} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x, Z = z\right).$$

Such a decomposition holds for any $j = 1, \ldots, m$. When the conditioning values (x, z) differ from the values (x_j, z_j) in the structural function, an application of Theorem 2.1 shows that the left hand side of this display represents a counterfactual conditional distribution, illustrating the connection between response types and counterfactual choices.

The following Theorem shows that, in order to rationalize a given collection of counterfactual conditional distribution under our assumptions, for each fixed θ it is both necessary and sufficient to construct a probability measure on sets of the form $\mathcal{U}(s,\theta)$ from (3.2) satisfying the constraints of Theorem 2.1. In the statement of the Theorem we redefine $\gamma : \mathbb{N} \to \mathbb{N}$ to denote the index of the point in $\{(x_1, z_1), \ldots, (x_m, z_m)\}$ assigned under counterfactual γ and we set $S_{\gamma(j)} := \{s \in \{0,1\}^m : s_{\gamma(j)} = 1\}$ (the analog of S_j from (3.3)).

Theorem 3.1. *Suppose Assumptions 2.1 and 2.2 hold. Fix some $\theta \in \Theta$ and consider the collection of sets:*

$$\mathcal{A}(\theta) := \{\mathcal{U}(s,\theta) : s \in \{0,1\}^m\}. \tag{3.4}$$

Then for any collection of counterfactual conditional distributions $P_{Y_\gamma|Y,X,Z}$, there exists a collection of Borel conditional probability measures $P_{U|Y,X,Z}$ satisfying (2.5) with $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ if and only if there exists a collection $P_{U|Y,X,Z}$ of probability measures on the sets in $\mathcal{A}(\theta)$ satisfying:

$$\sum_{s \in S_j} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = 1, X = x_j, Z = z_j\right) = 1, \tag{3.5}$$

$$\sum_{s \in S_j^c} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = 0, X = x_j, Z = z_j\right) = 1, \tag{3.6}$$

$$\sum_{s \in S_{\gamma(j)}} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x_j, Z = z_j\right) = P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j\right), \tag{3.7}$$

for all $y \in \{0,1\}$ and $j \in \{1, \ldots, m\}$ assigned positive probability.

Theorem 3.1 reduces our infinite dimensional existence problem to a finite dimensional existence problem. Indeed, the constraints in (3.5) and (3.6) are linear constraints on a now finite dimensional probability vector with typical element $P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x, Z = z\right)$. Note that this result relies crucially on the finiteness of $\mathcal{X}$ and $\mathcal{Z}$. Our proof of Theorem 3.1 appears to be new, and thus may be of separate interest.

[12] It is useful to note that the sets $\{S_j\}_{j=1}^m$ are not disjoint; indeed, it is easy to show that $S_j \cap S_k \neq \varnothing$ and $S_j \cap S_k^c \neq \varnothing$ for every $j \neq k$.

However, there is a close connection between Theorem 3.1 and the bounding approach based on Artstein's inequalities (e.g. Chesher and Rosen (2017)) and optimal transportation (e.g. Galichon and Henry (2011)).[13] Importantly, the finite number of linear constraints from Theorem 3.1 leads naturally to the optimization formulation of bounds on counterfactual distributions considered in the next subsection.

3.1 Optimization Formulation

We suppose throughout this subsection that our objective is to bound the counterfactual probability:

$$P_{Y_\gamma|Y,X,Z}(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j), \tag{3.8}$$

for some $j \in \{1, \ldots, m\}$. However, all the results in this section are immediately applicable to the case when we wish to bound some linear function of these counterfactual probabilities. Recall that Theorem 3.1 implies our counterfactual object of interest can be rewritten as:

$$P_{Y_\gamma|Y,X,Z}(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j) = \sum_{s \in S_{\gamma(j)}} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x_j, Z = z_j\right),$$

where $\gamma(j)$ is the index in $\{1, \ldots, m\}$ assigned to j under counterfactual γ. To progress further, let us define the parameter:

$$\pi(y, x, z, s, \theta) = P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x, Z = z\right).$$

For the sake of notation it is also useful to define the following parameter vectors:

$$\pi(y, s, \theta) := \begin{bmatrix} \pi(y, x_1, z_1, s, \theta) & \pi(y, x_1, z_2, s, \theta) & \ldots & \pi(y, x_{m_x}, z_{m_z}, s, \theta) \end{bmatrix}^\top,$$

$$\pi(y, \theta) := \begin{bmatrix} \pi(y, s_1, \theta)^\top & \pi(y, s_2, \theta)^\top & \ldots & \pi(y, s_{2^m}, \theta)^\top \end{bmatrix}^\top, \qquad \pi(\theta) := \begin{bmatrix} \pi(0, \theta)^\top & \pi(1, \theta)^\top \end{bmatrix}^\top.$$

The vector of parameters $\pi(\theta)$ represents the variable over which we optimize in our result ahead. Now let $d_\pi = 2m2^m$ denote the dimension of $\pi(\theta)$. Without loss of generality, we suppose that each (y, x, z) is assigned positive probability by the observed distribution. From conditions (3.5) and (3.6) in Theorem 3.1, we have the constraints:

$$\sum_{s \in S_j} \pi(1, x_j, z_j, s, \theta) = 1, \qquad \sum_{s \in S_j^c} \pi(0, x_j, z_j, s, \theta) = 1, \tag{3.9}$$

[13] In a previous version of this paper (Gu and Russell (2021)) we show that Theorem 3.1 is equivalent to a characterization based on Artstein's inequalities after conditioning on the value of the endogenous variables. This conditioning allows us to obtain a much smaller number of *equality* constraints when compared to the full set of unconditional constraints arising from Artstein's inequalities. It is also well known that Artstein's inequalities are equivalent to the existence of a certain zero-cost optimal transport problem (see Galichon (2016)).

for $j = 1, \ldots, m$. Finally, we require the nonnegativity and "adding-up" constraints:

$$\pi(y, x_j, z_j, s, \theta) \in \begin{cases} \{0\}, & \text{if } \operatorname{int}(\mathcal{U}(s, \theta)) = \varnothing, \\ [0, 1], & \text{otherwise,} \end{cases} \tag{3.10}$$

for all $y \in \{0, 1\}$ and $j = 1, \ldots, m$ and $s \in \{0, 1\}^m$, and:

$$\sum_{s \in \{0,1\}^m} \pi(y, x_j, z_j, s, \theta) = 1, \tag{3.11}$$

for all $y \in \{0, 1\}$ and $j = 1, \ldots, m$. Note that the researcher must determine which sets $\mathcal{U}(s, \theta)$ have nonempty interior in order to impose the constraint (3.10). We will return to this point in the next subsection. We are now ready to state one of the main results for this section.

Theorem 3.2. *Under Assumptions 2.1 and 2.2, the identified set for the counterfactual conditional probability $P_{Y_\gamma|Y,X,Z}(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j)$ is given by:*

$$\bigcup_{\theta \in \Theta} [\pi_{\ell b}(y, x_j, z_j, \theta), \pi_{ub}(y, x_j, z_j, \theta)], \tag{3.12}$$

where $\pi_{\ell b}(y, x_j, z_j, \theta)$ and $\pi_{ub}(y, x_j, z_j, \theta)$ are determined by the optimization problems:

$$\pi_{\ell b}(y, x_j, z_j, \theta) := \min_{\pi(\theta) \in \mathbb{R}^{d_\pi}} \sum_{s \in S_{\gamma(j)}} \pi(y, x_j, z_j, s, \theta), \text{ subject to (3.9), (3.10), and (3.11),} \tag{3.13}$$

$$\pi_{ub}(y, x_j, z_j, \theta) := \max_{\pi(\theta) \in \mathbb{R}^{d_\pi}} \sum_{s \in S_{\gamma(j)}} \pi(y, x_j, z_j, s, \theta), \text{ subject to (3.9), (3.10), and (3.11).} \tag{3.14}$$

In one direction, Theorem 3.2 implies that any counterfactual conditional probability of the form (3.8) belonging to the identified set can be written as:

$$P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j\right) = \sum_{s \in S_{\gamma(j)}} \pi(y, x_j, z_j, \theta, s),$$

for some θ and some vector $\pi(\theta)$ satisfying the constraints (3.9), (3.10), and (3.11). In the opposite direction, the Theorem implies that if for some θ the vector $\pi(\theta)$ satisfies the constraints (3.9), (3.10), and (3.11) then the conditional probability measure on $\mathcal{U}$ represented by $\pi(\theta)$ can be extended to a (not necessarily unique) Borel probability measure on all of $\mathfrak{B}(\mathcal{U})$ that satisfies the conditions of Theorem 2.1. This result can be easily modified to bound any linear function of counterfactual conditional distributions by simply modifying the objective function in Theorem 3.2. We will make use of this fact in the application section.

After determining which of the sets $\mathcal{U}(s, \theta)$ are empty, all the constraints in (3.13) and (3.14) can be written as linear equality/inequality constraints, so that the optimization problems in (3.13) and (3.14) are linear programming problems. This is very beneficial, since linear programs can be efficiently solved even in cases with thousands of parameters and constraints. It is also interesting to note that the proofs for Theorems 3.1 and 3.2 do not require linearity of the index function in U from Assumption 2.1 (although this assumption will be used heavily starting in the next subsection). This implies that Theorem 3.2 can be used to bound counterfactual parameters for completely nonparametric and nonseparable models of the form:

$$Y = \mathbb{1}\{\varphi(X, Z, U, \theta) \geq 0\}, \tag{3.15}$$

without imposing any restrictions on the index function. Without any constraints on the function φ it is always possible to construct a function φ such that all regions $\mathcal{U}(s, \theta)$ will have non-empty interior, implying that that there are 2^m response types. In this case, constraint (3.10) reduces to simply imposing that all probabilities are bounded between zero and one, and the rest of Theorem 3.2 remains unchanged. We will illustrate our procedure using the model (3.15) in the application section.

In the general case, elements of $\pi(\theta)$ corresponding to sets $\mathcal{U}(s, \theta)$ that are empty can be removed from the parameter vector $\pi(\theta)$ without altering the optimal solutions to the linear programs in (3.13) and (3.14). This allows for further reduction of the dimension of these linear programs. Although the number of sets $\mathcal{U}(s, \theta)$ appear to grow exponentially in m, in the subsections ahead we show that under linearity of the index function in U the number of sets $\mathcal{U}(s, \theta)$ that have non-empty interior grows at a rate that is polynomial in m, substantially reducing the computational burden. Finally, although Theorem 3.2 is an identification result we discuss estimators and inference procedures for bounds of the form in Theorem 3.2 in Section 5 when we introduce our application. In the next subsections we discuss an efficient algorithm for determining which of the sets $\mathcal{U}(s, \theta)$ have non-empty interior and discuss how to practically take the union in (3.12).

3.2 Hyperplane Arrangements and Cell Enumeration

As mentioned above, we have not yet used the fact that the index function $\varphi(X, Z, U, \theta)$ is restricted to be linear in $U \in \mathcal{U}$ under Assumption 2.1. However, linearity of the index function imposes restrictions on the model by limiting the number of sets $\mathcal{U}(s, \theta)$ that can be assigned positive probability. These restrictions enter the optimization problems in Theorem 3.2 implicity through the constraint (3.10). Reducing the number of sets that can be assigned positive probability imposes additional constraints in the optimization problems of Theorem 3.2 that help to tighten the identified set, and it can also reduce computational time needed to solve the bounding problems in Theorem 3.2 by reducing the dimension of the vector $\pi(\theta)$.

The assumption of linearity in the latent variables implies that certain response types must have zero probability. Constraining sets of the form $\mathcal{U}(s, \theta)$ to be assigned zero probability is referred to as *eliminating response types*. Response types corresponding to sets $\mathcal{U}(s, \theta)$ that survive elimination are called *admissible*, and response types corresponding to sets $\mathcal{U}(s, \theta)$ that are eliminated are called *inadmissible*. Since each response type is characterized by a particular menu of counterfactual responses, framing functional form assumptions in terms of the elimination of particular response types helps to provide some interpretation to these assumptions.

The following simple example shows how some of the sets $\mathcal{U}(s, \theta)$ can be empty under Assumption 2.1.

Example 1. *Suppose we have a variable $X \in \{0.5, 1, 2\}$ and latent variables $U \in \mathbb{R}^2$. That is, suppose there are no variables Z and no fixed coefficients θ. Then the structural function from (2.1) can be written as $\varphi(X, U)$ and the binary response vector $r(u, \theta)$ can be written as $r(u)$, where:*

$$r(u) = \begin{bmatrix} \mathbb{1}\{\varphi(0.5, u) \geq 0\} \\ \mathbb{1}\{\varphi(1, u) \geq 0\} \\ \mathbb{1}\{\varphi(2, u) \geq 0\} \end{bmatrix}.$$

Without any additional restrictions there is a total of $2^{|\mathcal{X}|} = 8$ possible response types.[14] *That is, $r(U) \in \{s_1, \ldots, s_8\}$, where:*

$$s_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad s_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad s_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad s_4 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad s_5 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad s_6 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad s_7 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad s_8 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Conclude that without any additional restrictions, all sets of the form $\mathcal{U}(s, \theta)$ for $s \in \{0, 1\}^3$ can be assigned positive probability by the optimization problems in Theorem 3.2. Now suppose we entertain a linear functional form restriction. In particular, suppose that Assumption 2.1 holds and that the structural function from (2.1) can be written as:

$$\varphi(X, U) = XU_1 - U_2.$$

Then the binary response vector $r(u)$ is given by:

$$r(u) = \begin{bmatrix} \mathbb{1}\{u_1 \geq 2u_2\} \\ \mathbb{1}\{u_1 \geq u_2\} \\ \mathbb{1}\{2u_1 \geq u_2\} \end{bmatrix}.$$

As is illustrated in Figure 1, under the assumption that the index function is linear in latent variables only 6 *response types are admissible. In particular, response types corresponding to binary vectors s_3 and s_6 are not possible under the linearity assumption. Thus, under Assumption 2.1 a distribution of latent variables is admissible in this example only if it assigns probability zero to the sets:*

$$\mathcal{U}(\theta, s_3) = \{u \in \mathcal{U} : r(u) = s_3\},$$
$$\mathcal{U}(\theta, s_6) = \{u \in \mathcal{U} : r(u) = s_6\}.$$

These additional constraints must be imposed in our optimization problems from Theorem 3.2.

This example shows that imposing linearity of φ in latent variables implies that certain sets of the form $\mathcal{U}(s, \theta)$ may be empty for some binary vectors $s \in \{0, 1\}^m$. In the general case, it can be shown that when φ is restricted to be linear in U, there is an upper bound on the number of non-empty sets $\mathcal{U}(s, \theta)$ that grows

[14] For example, if $U = (U_1, U_2)$ take $\varphi(X, U) = \sin(U_1 X + U_2)$ and fix $U_2 = 0$. Then it is straightforward to find eight values of the frequency parameter $U_1 \in [-1, 1]$ to rationalize each of the 8 response types.

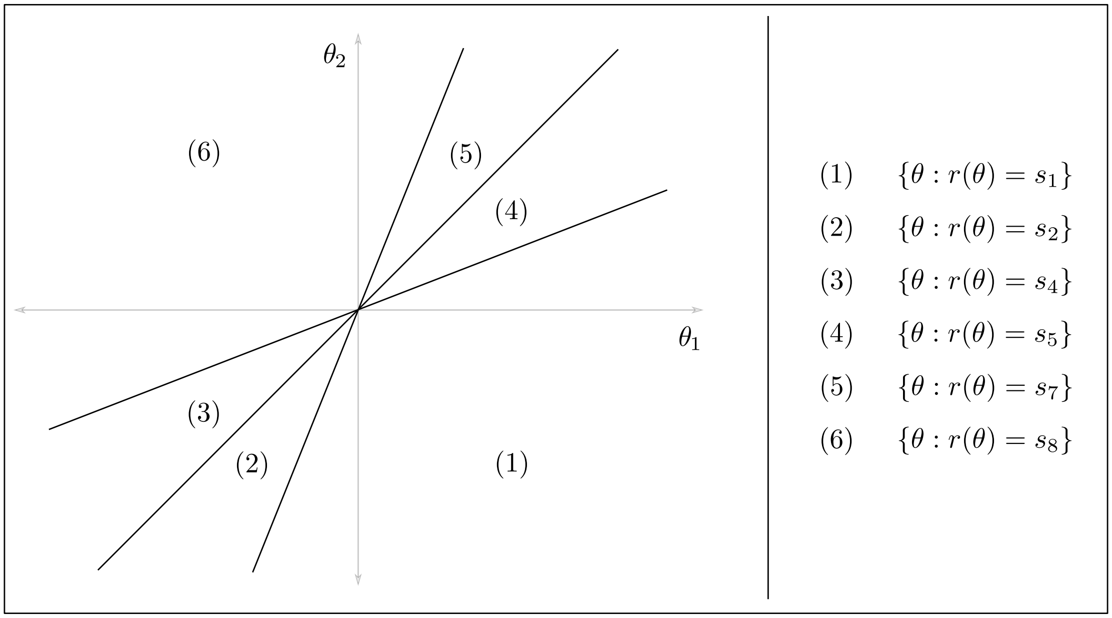

θ_2
(6)
(5)
(4)
θ_1
(3)
(2)
(1)
(1) $\{\theta : r(\theta) = s_1\}$
(2) $\{\theta : r(\theta) = s_2\}$
(3) $\{\theta : r(\theta) = s_4\}$
(4) $\{\theta : r(\theta) = s_5\}$
(5) $\{\theta : r(\theta) = s_7\}$
(6) $\{\theta : r(\theta) = s_8\}$

Figure 1: A figure corresponding to Example 1 illustrating the partition of the latent variable space according to response types in the case when the index function is linear. Without functional form restrictions, Example 1 shows 8 response types are possible; however, then the index function is linear in latent variables there are only 6 possible response types, as illustrated in the figure. In particular, the response types corresponding to binary vectors s_3 and s_6 from Example 1 are not possible.

at a rate that is polynomial in m (rather than exponential in m, which is the case when φ is unrestricted).

Proposition 3.1. *Suppose that Assumption 2.1 is satisfied. Then for each fixed $\theta \in \Theta$, there are at most $\sum_{j=0}^{d_u} \binom{m}{j}$ admissible response types.*

This result is implied by results in the literature on combinatorial geometry. In particular, linearity of the function $\varphi(\cdot, U)$ means that for each instance of (x, z, θ) the function $\varphi(x, z, u, \theta)$ defines a hyperplane in d_u-dimensional space. In the case when the vectors defining these hyperplanes are in *general position* the upper bound in Proposition 3.1 is obtained.[15] This latter result was initially proven by Buck (1943).

Let us define the collection of binary vectors S_φ to be those vectors $s \in \{0,1\}^m$ corresponding to admissible response types under Assumption 2.1. To impose linearity in the latent variables we must determine which sets $\mathcal{U}(s, \theta)$ have non-empty interior, and then ensure that any distribution of the latent variables under consideration when bounding counterfactual conditional distributions assigns zero probability to these sets. To practically implement our optimization problems we require a method of enumerating all admissible response types represented by the binary vectors in S_φ, and to compute the collection S_φ we propose to use the hyperplane arrangement algorithm of Gu and Koenker (2020).

When the index function φ is linear in U, for each fixed θ and $s \in \{0,1\}^m$ the set $\mathcal{U}(s, \theta)$ is a convex polyhedron formed by the intersection of halfspaces whose boundaries are hyperplanes of the form $\{u \in$

[15] A collection of m hyperplanes in $d-$dimensional space are considered to be in general position when if any collection of k out of the m hyperplanes intersect in a $d-k$ dimensional space for $1 < k \leq d$, and any collection of k out of m hyperplanes has an empty intersection for $k > d$.

$\mathcal{U} : \varphi(x, z, u, \theta) = 0\}$. Under Assumption 2.1 there are at most m such hyperplanes. The hyperplane arrangement algorithm of Gu and Koenker (2020) accepts these m hyperplanes as an input, and outputs the binary vectors s corresponding to the sets $\mathcal{U}(s, \theta)$ that have non-empty interior, as well as a point from each of these sets. In low dimensional space, it is relatively easy to determine the sets with non-empty interior formed by the intersection of halfspaces (see Figure 1, for instance). However, as the dimension of the space increases it becomes challenging to enumerate all of these sets. Avis and Fukuda (1996) were the first to provide an enumeration algorithm that runs in a time proportional to the maximum number of sets with non-empty interior. Improvements to this algorithm were made by Sleumer (1999) and Rada and Cerny (2018). The algorithm of Gu and Koenker (2020) is most closely related to the latter paper, and was developed for the problem of nonparametric maximum likelihood in a linear random coefficient model. It runs in a time proportional to $O(m^{d_u})$.

To understand the algorithm, note that for each $s \in \{0, 1\}^m$ and fixed θ, we can verify using a linear program whether there exists a point in the space of $\mathcal{U}$ that lies interior to the set $\mathcal{U}(s, \theta)$. Indeed, consider the following linear programming problem:

$$\max_{u,\varepsilon} \ \varepsilon \qquad \text{s.t.} \qquad (2s_j - 1)\varphi(x_j, z_j, u, \theta) \geq \varepsilon, \quad j = 1, \ldots, m, \tag{3.16}$$

where s_j is the j^{th} element of our fixed binary vector s, and where here we have an index function $\varphi(x, z, u, \theta)$ that is linear in u. If ε^* and u^* are the optimal values of the program (3.16) (provided that it is feasible), then an optimal value $\varepsilon^* > 0$ indicates that u^* is an interior point to the polyhedron $\mathcal{U}(s, \theta)$. However, since the linear program (3.16) must be solved for each $s \in \{0, 1\}^m$, checking whether each $\mathcal{U}(s, \theta)$ admits an interior point requires solving 2^m linear programs, despite the fact that we know the number of non-empty subsets $\mathcal{U}(s, \theta)$ is polynomial in m.

To address this issue, the algorithm proposed in Gu and Koenker (2020) builds upon the algorithm in Rada and Cerny (2018). The idea is to add one hyperplane at a time, and to enumerate all feasible response types after adding each new hyperplane. At step k they start with a collection of $k - 1$ hyperplanes from the previous steps, as well as all existing response types found up to step $k - 1$. They then introduce a new hyperplane into the arrangement of hyperplanes, and determine all newly created response types by solving a collection of linear programs. The algorithm of Rada and Cerny (2018) requires solving a linear programming problem for all the existing cells at each iteration, which amounts to solving $O(m^{d_u+1})$ such problems. When m is large, which is typically the case in practice, this can become costly. Gu and Koenker (2020) observed that when a new hyperplane is added the only new cells are those that are created when the existing cells are crossed by the last hyperplane. By efficiently locating those crossed cells, the algorithm reduced the number of linear programming problems to be solved by a magnitude of m. The algorithm in Gu and Koenker (2020) is available in the `R` package `RCBR`.

In summary, the hyperplane arrangement algorithm can be used as a pre-processing step under Assumption 2.1 to determine which sets $\mathcal{U}(s, \theta)$ have non-empty interior in a given application. Eliminating the inadmissible sets $\mathcal{U}(s,\theta)$ can also dramatically reduce the dimension of the parameter vector $\pi(\theta)$ in the bounding optimization problems. In particular, under Assumption 2.1 we need only consider a parameter vector $\pi(\theta)$ with typical element $\pi(y,x,z,\theta,s)$ defined only for s corresponding to subsets $\mathcal{U}(s,\theta)$ with non-empty interior. The dimension of the revised parameter vector $\pi(\theta)$ constructed in this way is always upper-bounded by a polynomial in m under Assumption 2.1. In the next subsection we show how the assumption of linearity in parameters $\theta \in \Theta$ can be combined with the hyperplane arrangement algorithm to dramatically simplify the bounding procedure suggested by Theorem 3.2.

3.3 Profiling Under Linearity in the Fixed Coefficients

To construct sharp bounds on counterfactual probabilities using Theorem 3.2 requires evaluating the linear programs (3.13) and (3.14) at all values of $\theta \in \Theta$ in the parameter space. In practice this procedure is clearly infeasible, and instead the identified set must be constructed using Theorem 3.2 by establishing a grid over the parameter space Θ, determining which of the sets $\mathcal{U}(s,\theta)$ have non-empty interior at each value of θ in the grid, and then solving the optimization problems (3.13) and (3.14) for each value of θ in the grid. The following proposition demonstrates that, theoretically speaking, the researcher need only repeat the procedure just described for *finitely* many values of θ.

Proposition 3.2. *Suppose that Assumptions 2.1 and 2.2 hold. Then there exists a (not necessarily unique) finite subset $\Theta' \subset \Theta$ such that:*

$$\left\{\overline{\pi} \in \mathbb{R}^{d_\pi} : \exists \theta \in \Theta \text{ s.t. } \pi(\theta) \text{ satisfies } (3.9), (3.10), (3.11), \text{ and } \overline{\pi} = \pi(\theta)\right\}$$
$$= \left\{\overline{\pi} \in \mathbb{R}^{d_\pi} : \exists \theta \in \Theta' \text{ s.t. } \pi(\theta) \text{ satisfies } (3.9), (3.10), (3.11), \text{ and } \overline{\pi} = \pi(\theta)\right\}.$$

We will call the points in the set Θ' the *representative points*, although it is important to keep in mind that these points are generally not unique. Assuming the representative points can be determined by the researcher, Proposition 3.2 immediately implies that the union over $\theta \in \Theta$ in (3.12) can be replaced with a union over $\theta \in \Theta'$. That is, the linear programs in (3.13) and (3.14) need only be solved at the representative points. Proposition 3.2 also implies that the identified set for counterfactual conditional distributions in Theorem 3.2 will always be a closed (but possibly disconnected) set. Unfortunately, when the researcher cannot determine the representative points Proposition 3.2 has limited practical value. In these cases, griding over the parameter space will often at best lead to an inner approximation to the identified set, and may be computationally prohibitive.

In the case when φ is linear in parameters we provide a polynomial-time algorithm for finding a collection of representative points. To introduce our approach, note that under the assumption that φ is linear in (U,θ), for each fixed $\theta \in \Theta$ the sets of the form $\mathcal{U}(s,\theta)$ define a unique partition of the space $\mathcal{U}$ into sets whose boundaries are defined by m hyperplanes. Let us define:

$$\mathcal{S}(\theta) := \{s \in \{0,1\}^m : \text{int}(\mathcal{U}(s,\theta)) \neq \varnothing\}.$$

Then $\mathcal{S}(\theta)$ denotes the set of all vectors $s \in \{0,1\}^m$ that are inducible by our arrangement of m hyperplanes. Now recall that functional form assumptions impose restrictions in the bounding optimization problems by restricting the number of sets $\mathcal{U}(s,\theta)$ with non-empty interior. For any two values of $\theta, \theta' \in \Theta$ with $\theta \neq \theta'$, if $\mathcal{S}(\theta) = \mathcal{S}(\theta')$ then the linear programming problems in Theorem 3.2 at θ and θ' are identical, since they have an identical set of constraints. The points θ and θ' are thus equivalent in the sense that we only need to solve the linear programming problems for one of them. Extending this idea, we can define an equivalence class by the set of all $\theta \in \Theta$ delivering the same collection $\mathcal{S}(\theta)$. We then only need to solve the linear programming problems at one value of θ belonging to each equivalence class. These values of θ selected from each equivalence class are exactly what we call representative points.

To see how to find the representative points, let us partition $U := (U_x, U_z, \varepsilon)$, $\theta = (\theta_x, \theta_z)$, $x = (x_r, x_f)$ and $z = (z_r, z_f)$, and for a binary vector $s \in \{0,1\}^m$ let us define the set:

$$\mathcal{R}(s) := \left\{ (u,\theta) : \begin{bmatrix} \mathbb{1}\{u_x x_{r1} + u_z z_{r1} + \theta_x x_{f1} + \theta_z z_{f1} \geq \varepsilon\} \\ \mathbb{1}\{u_x x_{r2} + u_z z_{r2} + \theta_x x_{f2} + \theta_z z_{f2} \geq \varepsilon\} \\ \vdots \\ \mathbb{1}\{u_x x_{rm} + u_z z_{rm} + \theta_x x_{fm} + \theta_z z_{fm} \geq \varepsilon\} \end{bmatrix} = s \right\}. \tag{3.17}$$

These sets form a unique partition of the space (u,θ) defined by m hyperplanes of the form:

$$u_x x_{ri} + u_z z_{ri} + \theta_x x_{fi} + \theta_z z_{fi} = \varepsilon. \tag{3.18}$$

The basic idea behind our strategy to find representative points is to first project the sets of the form $\mathcal{R}(s)$ onto the parameter space Θ. Note that the projection of a set $\mathcal{R}(s)$ onto the parameter space Θ delivers the set of all θ consistent with the binary vector s for some value of u. After taking the intersection of all such projections, each set in the resulting collection corresponds exactly to an equivalence class discussed above. An arbitrary value of θ taken from such a set is a representative point. The most challenging part of this approach is to find a tractable characterization of the projections of $\mathcal{R}(s)$ on the parameter space Θ.

Let us define $\mathcal{S}_p$ as the collection of all binary vectors $s \in \{0,1\}^m$ corresponding to the sets in $\mathcal{R}(s)$ with non-empty interior. The first step of our procedure to find the representative points is to determine the binary vectors in $\mathcal{S}_p$. This can be done by running the hyperplane arrangement algorithm of Gu and Koenker (2020) on the collection of hyperplanes of the form (3.18) defined on $\mathcal{U} \times \Theta$ (i.e. as if we were treating θ as a latent variable). Note that the assumption of linearity of φ in (U,θ) restricts the number of sets in the collection $\mathcal{S}_p$ to be polynomial in m.[16]

Next, let us define $w_{ri} := (x_{ri}, z_{ri}, -1)$ and $w_{fi} := (x_{fi}, z_{fi})$, where w_{ri} has dimension d_r and w_{fi} has

[16] Note that in this context, all the hyperplanes of the form (3.18) can be viewed as hyperplanes through the origin in $\mathcal{U} \times \Theta$. In this case, the upper bound on the number of cells formed by this collection of hyperplanes is of smaller order than that presented in Proposition 3.1. Cover (1965) shows the upper bound is given by:

$$C(m, d_u) := 2 \sum_{j=0}^{d_u - 1} \binom{m-1}{j}.$$

dimension d_f. Then each of the hyperplanes of the form (3.18) can be written as $w_{ri}u + w_{fi}\theta = 0$. Stacking these hyperplanes into matrix form we have $W_r u_r + W_f \theta = 0$, where W_r is $m \times d_r$ and W_f is $m \times d_f$. Now each set of the form (3.17) is a polyhedral cone in $\mathbb{R}^{d_x + d_z + 1}$ and can be uniquely identified by a sign vector $2s - 1$ with values in $\{-1, 1\}^m$. Fix any $s \in \mathcal{S}_p$, and let $D(s) = \text{diag}(2s - 1)$ denote the $m \times m$ diagonal matrix with the sign vector $2s - 1$ along its main diagonal. Furthermore, define $W_r(s) := D(s)W_r$ and $W_f(s) := D(s)W_f$. Then the set $\mathcal{R}(s)$ from (3.17) can be conveniently rewritten as:

$$\mathcal{R}(s) := \{(u, \theta) : W_r(s)u + W_f(s)\theta \geq 0\}.$$

Note that the row dimension of $W_r(s)$ and $W_f(s)$ is m, which can be large if the support $\mathcal{X} \times \mathcal{Z}$ contains many elements. Thus, in practice it is useful to first remove redundant inequalities among those that define $\mathcal{R}(s)$ before proceeding to the next step. Elimination of redundant inequalities from this system can be achieved in polynomial time with a sequence of linear programs, and the resulting set of nonredundant inequalities that define the polyhedral cone $\mathcal{R}(s)$ is typically much smaller than m.[17]

From here on we assume the matrices $W_r(s)$ and $W_f(s)$ only include rows corresponding to nonredundant constraints, and we denote their row dimension as $m(s)$. Now consider the set:

$$\Theta(s) := \{\theta \in \Theta : \exists u \in \mathcal{U} \text{ s.t. } W_r(s)u + W_f(s)\theta \geq 0\}. \tag{3.19}$$

In other words, the set $\Theta(s)$ is precisely the projection of the polyhedral cone $\mathcal{R}(s)$ on the parameter space Θ. The objective is to show that the set $\Theta(s)$ can be defined only in terms of linear inequality constraints in θ. In other words, we would like to "eliminate" the latent variables U from the system of inequalities in (3.19). A natural method of doing so is to use Fourier-Motzkin elimination.[18] Recall that the Fourier-Motzkin algorithm eliminates variables from a system of linear inequalities by taking linear combinations of the inequalities in the system. In particular, Fourier-Motzkin elimination can be viewed as applying a sequence of matrix operators $M_1, M_2, \ldots, M_{d_r}$ to the system of inequalities in (3.19), where the matrix $M_k M_{k-1} \ldots M_1$ eliminates the first k elements of the vector U from the inequalities. Let us denote $M_r^* = M_{d_r} M_{d_r - 1} \ldots M_1$. Then as a result of Fourier-Motzkin elimination we would have the equivalent system:

$$\Theta(s) := \{\theta \in \Theta : M_r^* W_f(s)\theta \geq 0\}, \tag{3.20}$$

since $M_r^* W_r(s) = 0$ by construction of M_r^*. The set in (3.20) then gives us inequality constraints only in terms of θ that define the projection of $\mathcal{R}(s)$ on Θ.

[17] In particular, not all the hyperplanes that define $\mathcal{R}(s)$ are relevant, in the sense that some of them are implied by the rest of the inequalities in the system. Removing these redundant inequalities does not change the cone $\mathcal{R}(s)$. We can remove them before continuing to the projection step of our procedure by conducting a redundancy test. For example, suppose we have system of $j + 1$ inequalities of the form $Ax \leq b$ and $s^\top x \leq t$. Then to check whether the last inequality is binding (and thus nonredundant), we can solve the linear programming problem $f^* = \max s^\top x$ s.t. $Ax \leq b, s^\top x \leq t + 1$. The inequality $s^\top x \leq t$ is redundant if and only if $f^* \leq t$. To eliminate all redundant inequalities from a system of m inequalities results in solving m linear programs; hence, it can be computed in polynomial time. There are a few strategies to speed up the removal of redundant inequalities, as discussed in Section 2.21 in Fukuda (2014). We use the implementation in the package `Rcdd` with the function `redundant`.

[18] The idea of using Fourier-Motzkin elimination to determine the inequality constraints defining projected regions in partial identification was also explored in Section 8.2 of Chesher and Rosen (2019).

While it is possible to use Fourier-Motzkin elimination to eliminate the latent variables U, the number of rows in the matrix M_r^* can be prohibitively large, even when the number of nonredundant inequalities defining the set (3.20) is small. To ensure feasibility of our method of projection, we must thus search for a procedure that eliminates redundant inequalities from (3.20) and results in a simpler characterization of $\Theta(s)$ than the one provided by Fourier-Motzkin elimination.[19] To this end, consider the following set:

$$\mathcal{C}(s) := \{c \in \mathbb{R}^{m(s)} : cW_r(s) = 0,\ c \geq 0\}, \tag{3.21}$$

where recall that $m(s)$ is the dimension of $W_r(s)$ and $W_f(s)$ after we've removed all the redundant inequalities. Since the rows of M_r^* have positive entries (by construction using the Fourier-Motzkin algorithm), they must belong to $\mathcal{C}(s)$. Thus we can conclude that:

$$\{\theta \in \Theta : cW_r(s)u + cW_f(s)\theta \geq 0,\ \forall c \in \mathcal{C}(s)\} \subseteq \Theta(s).$$

Furthermore, Kohler (1967) shows that the reverse inclusion holds; in particular, every vector in the collection $\mathcal{C}(s)$ can be written as a non-negative linear combination of the rows of M_r^*. We can thus conclude:

$$\Theta(s) = \{\theta \in \Theta : cW_r(s)u + cW_f(s)\theta \geq 0,\ \forall c \in \mathcal{C}(s)\}\,.$$

While at first glance this result is not immediately useful, the Minkowski-Weyl Theorem allows us to re-write the set $\mathcal{C}(s)$ as:

$$\mathcal{C}(s) = \left\{c \in \mathbb{R}^{m(s)} : c = R(s)a,\ \text{for some } a \geq 0\right\}, \tag{3.22}$$

where $R(s)$ is some matrix.[20] That is, every element belonging to the polyhedral cone $\mathcal{C}(s)$ can be written as a nonnegative linear combination of the columns of some matrix $R(s)$. It follows that if we could obtain the matrix $R(s)$ from (3.22), we could obtain the following representation of the projected set for θ from $\mathcal{R}(s)$:

$$\Theta(s) = \{\theta \in \Theta : H(s)\theta \geq 0\}\,, \tag{3.23}$$

where $H(s) = R(s)W_f(s)$. The matrix $R(s)$ is sometimes called the *generating matrix* of the polyhedral cone $\mathcal{C}(s)$. The Minkowski-Weyl Theorem essentially says that every polyhedral cone admits a generating matrix, and every generating matrix generates a polyhedral cone. The problem of finding the minimal generating matrix $R(s)$ (that is, the matrix $R(s)$ generating $\mathcal{C}(s)$ such that no proper submatrix of $R(s)$ generates $\mathcal{C}(s)$) is called the *extreme ray enumeration problem.* Note the minimal generating matrix is unique only up to

[19] Note that it is possible to first perform Fourier-Motzkin elimination, and then remove redundant inequalities from the system $M_r^* W_r(s)\theta \geq 0$ using a technique similar to that described in footnote 17. However, unlike Fourier-Motzkin elimination our alternate approach uses the double-description algorithm to avoid the generation of redundant inequalities altogether, and so is much more efficient.

[20] For a general convex polyhedral defined by $\Lambda = \{\lambda \in \mathbb{R}^d : A\lambda \leq b\}$, the Minkowski-Weyl Theorem states that every vector $\lambda \in \Lambda$ can be written as $\lambda = \lambda_1 + \lambda_2$, where $\lambda_1 \in \text{conv}\{v_1, \ldots, v_k\}$ and $\lambda_2 \in \text{cone}\{v_{k+1}, \ldots, v_n\}$. Here $v_1, \ldots, v_k$ are called vertices of Λ and $v_{k+1}, \ldots, v_n$ are the extreme rays of Λ. In the special case of $b = 0$, where all hyperplanes are through the origin, then Λ becomes a polyhedral cone and $k = 0$, so that $\Lambda = \text{cone}\{v_1, \ldots, v_n\}$. This latter case is what is relevant for us, and the columns of the matrix $R(s)$ are the collections of these extreme rays.

multiplication by a positive scalar.

The characterizations of the cone $\mathcal{C}(s)$ in (3.21) and (3.22) are called its H-representation and its V-representation, respectively. Converting from one representation of a convex polyhedron to another is called the *double description problem* in computational geometry, and is one of the most important problems in the field of polyhedral computation. One of the earliest double description algorithms proposed by Motzkin et al. (1953) is an incremental algorithm that computes in exponential time in the worst case. The idea is to start with a small subset of hyperplanes and an associated V-representation and continue to add hyperplanes while updating the set of extreme rays. However, the performance of the algorithm can be sensitive to the order that new hyperplanes are introduced. A more efficient variant of this procedure is proposed by Fukuda and Prodon (1995), and we use the `R` implementation in the package `Rcdd` by Geyer (2019). There are alternative nonincremental algorithms available for extreme ray enumeration; for instance, the reverse search algorithm by Avis and Fukuda (1996). However, in general there is no known efficient (polynomial-time) algorithm for general input, although the incremental double description algorithm is known to be efficient for degenerate polyhedrons (which arises very often when the hyperplanes are not in general position) and low dimensions (up to 10).[21] Avis et al. (1997) present a thorough comparison of these different algorithms.

After employing the double description algorithm the projection $\Theta(s)$ represented in (3.23) contains a minimal number of constraints defined only in terms of θ. Repeating the procedure described above for all $s \in \mathcal{S}_p$ then gives us a collection of sets $\Theta(s)$ representing the projections of $\mathcal{R}(s)$ onto the parameter space Θ. However, for different binary vectors $s \in \mathcal{S}_p$ the projected sets $\Theta(s)$ may not be disjoint. Thus, to get the representative points θ^* we consider the intersection of these cones across $s \in \mathcal{S}_p$. To do so, we stack all unique hyperplanes of the form $H(s)\theta = 0$ for all $s \in \mathcal{S}_p$ into a matrix H_p. The set of hyperplanes $H_p\theta = 0$ then define the boundaries of the sets formed by the intersection of the cones $\Theta(s)$. From here we can then easily collect the representative points from the resulting collection of sets defined by the hyperplanes $H_p\theta = 0$ by a final application of the hyperplane arrangement algorithm of Gu and Koenker (2020).

To summarize, our procedure to profile θ is based on the idea that there are only a finite number of representative points from Θ that need to be considered in the bounding optimization problems. Our proposed procedure to find these representative points is as follows:

(i) Determine the collection $\mathcal{S}_p$ of binary vectors $s \in \{0,1\}^m$ corresponding to the sets $\mathcal{R}(s)$ from (3.17) with non-empty interior by running the hyperplane arrangement algorithm of Gu and Koenker (2020) on the collection of hyperplanes of the form (3.18).

(ii) For each $s \in \mathcal{S}_p$:

(a) Set $D(s) = \text{diag}(2s - 1)$ and define $W_r(s) := D(s)W_r$ and $W_f(s) := D_f(s)W_f$. Now remove any

[21] For an incremental algorithm to be polynomial-time, the size of the intermediate rays in each incremental step needs to be polynomial in the input size. The difficulty involved with all known incremental algorithm in the literature is that the intermediate representation can be very large and leads the algorithm to be superpolynomial in the worst case. See further discussion in Bremner (1999).

redundant inequalities from the system of inequalities in the set:

$$\mathcal{R}(s) := \{(u,\theta) : W_r(s)u + W_f(s)\theta \geq 0\},$$

by solving a sequence of linear programs, as described in footnote 17.

(b) Compute the minimal generating matrix $R(s)$ for the polyhedral cone $C(s)$ using the double description algorithm of Fukuda and Prodon (1995), and set $H(s) = R(s)W_f(s)$. Then the projected set $\Theta(s)$ from (3.19) can be written:

$$\Theta(s) = \{\theta \in \Theta : H(s)\theta \geq 0\}.$$

(iii) Intersect the projected sets $\Theta(s)$ over all $s \in \mathcal{S}_p$. By stacking the matrices $H(s)$ over $s \in \mathcal{S}_p$ into the matrix H_p, the rows of the matrix H_p defines a set of hyperplanes that act as the boundaries of all sets defined by the intersection of the projected sets $\Theta(s)$.

(iv) Run the hyperplane arrangement algorithm of Gu and Koenker (2020) a final time on the collection of hyperplanes defined by the rows of H_p in order to collect representative points from each set.

The above discussion sheds light on how we can construct the identified set for θ. In particular, for some of these representative points the linear programming problems in our bounding procedure may have an empty feasible region, that is, there exists no valid conditional distribution of u that fulfils all constraints for that particular value of θ. In this case, these representative points—as well as all other values of θ that belong to the same sets—cannot be included in the identified set for the fixed coefficients Θ^*. Therefore, the identified set Θ^* naturally collects all sets whose representative points render a linear program with non-empty feasible region. Since the arrangement involves only hyperplanes through the origin, all sets take the form of a polyhedral cone, hence the identified set Θ^* is a union of polyhedral cones. This implies that the identified set Θ^* may not be connected, and for any $\theta \in \Theta^*$, we also have $\lambda\theta \in \Theta^*$ for all $\lambda \geq 0$. An appropriate normalization—for example, fixing $||\theta|| = 1$—leads to a bounded identified set Θ^*.

4 Additional Assumptions

The previous section outlines the mechanics of our main bounding procedure. In this section we show how to introduce additional independence and monotonicity assumptions. Independence assumptions are quite common in parametric binary response models and binary response models with endogenous regressors, although here we show how to impose various independence assumptions as a set of linear equality constraints in the optimization problems of Theorem 3.2. Finally, monotonicity assumptions appear in various forms in the literature on treatment effects, and our incorporation of monotonicity restrictions arising from choice theory makes substantial use of response types, resembling the approach of Heckman and Pinto (2018).

4.1 Independence Assumptions

In some cases the researcher may have access to a variable that is believed to be independent of the distribution of latent variables. If such a variable enters as an argument in the structural function, then intuitively such a variable induces variation in the observed conditional probabilities without affecting the distribution of latent variables. We refer to such variables as *exogenous covariates*. A similar intuition applies if the variable is independent of the distribution of latent variables, does not enter as an argument in the structural function, but has nontrivial dependence with the variables that do enter the structural function.[22] We refer to such variables as *instruments*. Any additional variation generated in the observed conditional probabilities by either exogenous covariates or instruments can be used to further pin down the distribution of latent variables.

We now distinguish between the random variables in X and Z by allowing the variables in the random vector Z to satisfy an independence assumption with the latent variables U.

Assumption 4.1 (Independence). *For all $A \in \mathfrak{B}(\mathcal{U})$ we have $P_{U|Z}(A \mid Z = z) = P_U(A)$, P_Z−a.s.*

The independence assumption restricts the econometric model by constraining the set of admissible latent variable distributions, and provides a crucial link between the conditional distributions of $U \mid Z = z$ across values of $z \in \mathcal{Z}$. When applied to our context, Assumption 4.1 nests the two kinds of independence constraints introduced above (i.e. exogenous covariates and instruments). Furthermore, it is without loss of generality that we continue to write the structural function φ as a function of Z, which helps us avoid unnecessary repetition by considering the two kinds of independence constraints separately. Also, even though Assumption 4.1 posits full independence between Z and the vector of latent variables U, the assumption can be easily modified for the case when a subvector of Z, say Z_1, is conditionally independent of U given some other subvector of Z, say Z_2. We suppress this case for simplicity, but we note that consideration of conditional independence does not have any significant impact on the results to come, and thus can be easily accommodated.

Definition B.1 in Appendix B.1 provides the extension of Definition 2.1 to the case when Assumption 4.1 also holds. Corollary B.1 in Appendix B.1 then provides the extension of Theorem 3.1 to the case when Assumption 4.1 also holds, and again allows us to reduce an infinite dimensional existence problem to a manageable finite dimensional existence problem. Intuitively, Corollary B.1 shows that every conditional probability measure $P_{U|Y,X,Z}$ defined on the sets $\mathcal{A}(\theta)$ from (3.4) satisfying the independence assumption can be extended to a probability measure on $\mathfrak{B}(\mathcal{U})$ that satisfies Assumption 4.1. This result can be used to show that Assumption 4.1 is observationally equivalent to imposing independence between Z and response types $r(U, \theta)$.

To extend the linear programming result of Theorem 3.2 it is straightforward to see that we must simply include the additional constraints from Corollary B.1. Without loss of generality we again assume that all

[22] Restricting an exogenous variable from entering the structural function is known as the *exclusion restriction* in the terminology of simultaneous equations.

values of (y, x, z) are assigned positive probability by the observed distribution. Then these constraints can be written in terms of the parameter vector $\pi(\theta)$ as:

$$\sum_{y\in\{0,1\}} \sum_{x\in\mathcal{X}} \pi(y, x, z_k, s, \theta) P(Y = y, X = x \mid Z = z_k)$$
$$= \sum_{y\in\{0,1\}} \sum_{x\in\mathcal{X}} \pi(y, x, z_{k+1}, s, \theta) P(Y = y, X = x \mid Z = z_{k+1}), \tag{4.1}$$

for $k = 1, \dots, m_z - 1$. The formal statement of the extension of Theorem 3.2 to the case when the constraints (4.1) are also imposed is provided by Corollary B.2 in Appendix B.1.

4.2 Monotonicity Assumptions

In this subsection we introduce monotonicity assumptions and demonstrate how monotonicity assumptions impose constraints on the bounding problem by effectively eliminating certain response types. To introduce our monotonicity assumptions, let $\mathcal{M} \subset \{1, \dots, m\} \times \{1, \dots, m\}$ denote any collection of pairs of integers (j, k), where $1 \leq j, k \leq m$.

Assumption 4.2 (Monotonicity)**.** *For each $\theta \in \Theta$ and each pair (j, k) in the set $\mathcal{M}$ we have $\varphi(x_j, z_j, \theta, u) \leq \varphi(x_k, z_k, \theta, u)$.*

This monotonicity assumption states that, when comparing two points (x_j, z_j) and (x_k, z_k), the value of the structural function can be ordered by the researcher. Note that if the order determined by the researcher's monotonicity assumption for the pair of points (x_j, z_j) and (x_k, z_k) is $\varphi(x_j, z_j, \theta, u) \leq \varphi(x_k, z_k, \theta, u)$ (for example), then the researcher automatically rules out response types with $\mathbb{1}\{\varphi(x_j, z_j, \theta, u) \geq 0\} > \mathbb{1}\{\varphi(x_k, z_k, \theta, u) \geq 0\}$. The following example illustrates how this idea leads to elimination of response types.

Example 2. *Suppose again that we have only a binary variable $X \in \{0, 1\}$ and latent variables U (i.e. no variables Z and no fixed coefficients θ). Then the structural function from (2.1) can be written as $\varphi(X, U)$ and the binary response vector $r(u, \theta)$ can be written as $r(u)$, where:*

$$r(u) = \begin{bmatrix} \mathbb{1}\{\varphi(0, u) \geq 0\} \\ \mathbb{1}\{\varphi(1, u) \geq 0\} \end{bmatrix}.$$

Note that there are only four response types; that is, $r(u) \in \{s_1, s_2, s_3, s_4\}$ where:

$$s_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad s_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad s_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad s_4 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Without any additional restrictions, all response types—and thus all sets of the form $\mathcal{U}(s, \theta)$ for $s \in \{0, 1\}^2$—can be assigned positive probability by the optimization problems in Theorem 3.2. Now suppose we entertain the monotonicity assumption $\varphi(0, u) \leq \varphi(1, u)$. Imposing this constraint clearly rules out the case when

$r(u) = s_2$, and thus the set $\mathcal{U}(\theta, s_2) = \{u : r(u) = s_2\}$ must be assigned probability zero in any solution to the optimization problems in Theorem 3.2. Constraining such sets to be assigned zero probability in these optimization problems reduces the size of the feasible region and thus potentially tightens the resulting bounds on counterfactual probabilities.

Monotonicity of the type entertained here has a number of precedents in the literature on treatment effects, and can be interpreted a few different ways. For example, when Y is interpreted as a treatment indicator, the type of monotonicity introduced here nests the monotonicity assumption from Angrist et al. (1996) required for identification of the *local average treatment effect*. Alternatively, when Y is the interpreted as the binary outcome after some (possibly endogenous) treatment X, our monotonicity assumption can be interpreted as a version of the *monotone treatment response* assumption introduced in Manski (1997) and also considered in Manski and Pepper (1998). Finally, similar monotonicity assumptions in triangular systems have been extensively explored by Heckman and Pinto (2018). In particular, Heckman and Pinto (2018) explore how choice theory can be used to impose monotonicity assumptions and to eliminate response types, and many of their insights are applicable here.

Following the insights from the example above, let us define the collection of binary vectors S_M to be those that respect the monotonicity relations from Assumption 4.2. Definition B.2 in Appendix B.2 provides the extension of Definition 2.1 to the case when Assumption 4.2 is also imposed. The extension of Theorem 3.1 to the case when Assumption 4.2 is imposed is provided by Corollary B.3 in Appendix B.2. To extend the results of Theorem 3.2 we must simply include the set of constraints imposed by Assumption 4.2 in our optimization problems. These constraints are provided in Corollary B.3, and can be written in terms of the parameter vector $\pi(\theta)$ as:

$$\sum_{s \in S_M^c} \pi(y, x_j, z_j, s, \theta) = 0, \tag{4.2}$$

for all $y \in \{0, 1\}$ and $j = 1, \ldots, m$ occurring with positive probability. Corollary B.4 in Appendix B.2 then shows the extension of Theorem 3.2 to the case when Assumption 4.2 is imposed using the constraints (4.2).

Combining all the results seen in this section, any combination of Assumption 2.1, Assumption 4.1 and Assumption 4.2 can be imposed on the optimization problems in Theorem 3.2 by simply adding the corresponding combination of constraints (4.1) and/or (4.2).

5 Application

In this section we apply our method to study the impact of private health insurance on an individual's decision to visit a doctor. In general, insurance markets are plagued by problems arising from asymmetric information between consumers and insurance providers (c.f. Rothschild and Stiglitz (1978)). For example, adverse selection occurs in the health insurance market when individuals have more information about their latent health determinants than the providers of health insurance. A robust prediction of the classical theory of asymmetric information is that those who are more likely to purchase insurance are also those who are more likely to experience the insured risk.[23] On the other hand, there has been little and mixed empirical evidence of adverse selection in health insurance markets (see Cardon and Hendel (2001) for a discussion). Others have suggested that those who purchase insurance may be more risk averse, and so less likely to engage in activities that might cause them to experience the insured risk. Evidence of this is found in Finkelstein and McGarry (2006), who demonstrate that wealthier and more cautious individuals are more likely to have long-term care insurance, but less likely to ever use their insurance. However, in many cases the opposite is equally plausible. For example, Bajari et al. (2014) explore the effect of moral hazard in health insurance markets, which occurs when those who purchase health insurance are more likely to experience the insured risk given that they no longer bear the full cost of health care.

Here we do not attempt to disentangle the effects of adverse selection, risk aversion, or moral hazard. Instead we compute various counterfactual parameters while remaining agnostic on the exact nature of the unobservables linking the health insurance and health care utilization decisions. We take the decision to visit a doctor as our binary outcome variable of interest, and we consider the individuals' private health insurance status to be an endogenous explanatory variable. This latter point is consistent with the idea that private insurance status may be dependent with individual-specific latent factors—most importantly, unobserved health determinants and attitudes towards risk—that influence an individual's propensity to visit a doctor. We use data from the 2010 wave of the Medical Expenditure Panel Survey (MEPS). This data has been analyzed by Han and Lee (2019). We focus on the same sub-sample they consider. In particular, we focus on the month of January 2010, consider only individuals between ages 25 and 64, and drop individuals who obtain either federal or state insurance in 2010 and individuals who are self-employed or unemployed. These restrictions leave us with a sample of 7555 individuals.

In all specifications X is a binary endogenous variable representing an individual's private insurance status, and we consider a binary health status variable (Z_1) and a binary marital status variable (Z_2) as regressors.[24] Finally, we use the number of employees working for the individual's firm (Z_3) as an instrument. This variable provides a measure of the size of a firm and has discrete support in the range $[1, 500]$, which we further discretize into 11 bins.[25] Using firm size as an instrument is consistent with the evidence that larger firms are more likely to provide health insurance benefits, but do not directly influence an individual's decision to visit a doctor.[26] The same instrument was used in Han and Lee (2019).

A possible concern with using firm size as an instrument is that risk averse individuals may be more

[23] The "insured risk" refers to the event for insurance was purchased. In our context, it is any event that would typically require a visit to the doctor.

[24] The MEPS data includes information on self-reported health status on a scale from $1 - 5$, and we regard values less than or equal to 2 as being "unhealthy."

[25] Variable Z_3 is supported on the range $[1, 500]$ which is clearly top-coded. We notice that there is bunching of observations at firm size rounded by five, which implies that some of the support of Z_3 has very few observations. In order to get reliable estimates of the conditional choice probabilities, we further discretize the firm size into 11 bins. The bins are respectively $[1, 5]$, $(5, 10]$, $(10, 20]$, $(20, 30]$, $(30, 40]$, $(40, 50]$, $(50, 60]$, $(60, 70]$, $(70, 100]$, $(100, 200]$ and $(200, 500]$.

[26] From Cardon and Hendel (2001) p.408: "Another observed symptom, consistent with the theoretical predictions, is that the uninsured tend to work for small employers. Large employers can overcome adverse selection by risk pooling."

likely to select into a job with a larger firm size. In an attempt to address this issue, we investigate a weaker independence assumption (which we call *relaxed independence*) that assumes the firm size Z_3 is conditionally independent of U given (Z_1, Z_2) only when Z_3 lies within a certain range. The main idea is that once we condition on a particular range of firm size, the remaining variation in firm size is independent of U conditional on (Z_1, Z_2). We consider four ranges, given by $(1, 10]$, $(10, 50]$, $(50, 100]$ and $(100, 500]$, and impose our conditional independence assumption for each range separately.

The first parameter we consider is the average treatment effect, defined as:

$$\begin{aligned}\mu_{ate} := &\sum_{(y,x,z)\in\{0,1\}\times\mathcal{X}\times\mathcal{Z}} P_{U|Y,X,Z}(\varphi(1, z, u, \theta) \geq 0 \mid Y = y, X = x, Z = z)P(Y = y, X = x, Z = z)\\ &- \sum_{(y,x,z)\in\{0,1\}\times\mathcal{X}\times\mathcal{Z}} P_{U|Y,X,Z}(\varphi(0, z, u, \theta) \geq 0 \mid Y = y, X = x, Z = z)P(Y = y, X = x, Z = z).\end{aligned}$$

This parameter provides the average causal effect of obtaining health insurance on the decision to visit a doctor. Near the end of this section we also consider bounds on counterfactual conditional choice probabilities. We construct our bounds under the following set of assumptions:

(A1) Only Assumptions 2.1 and 2.2.

(A2) (A1) and monotonicity (Assumption 4.2). The discussion below provides further details.

(A3) (A1) and independence between (Z_1, Z_2) and U (Assumption 4.1).

(A4) (A1), (A2) and (A3) together.

(A5) (A1) and independence between (Z_1, Z_2, Z_3) and U (Assumption 4.1).

(A6) (A1), (A2) and (A5) together.

(A7) (A1) and independence between (Z_1, Z_2) and U, and relaxed independence with Z_3 (Assumption 4.1).

(A8) (A1), (A2) and (A7) together.

Note that the general index function takes the form $\varphi(x, z_1, z_2, u, \theta)$, and when we say that the monotonicity assumption is imposed in (A2), we are in fact imposing:

$$\varphi(1, 0, z_2, u, \theta) \geq \varphi(0, 0, z_2, u, \theta),$$

for each $z_2 \in \{0, 1\}$. This implies that for an unhealthy individual, the propensity to visit a doctor when the person has private insurance is always weakly greater than without the insurance, regardless of marital status. Finally we consider three different models for the binary outcome variable Y:

$$Y = 1\{\varphi(X, Z, U) \geq 0\}, \tag{M1}$$

$$Y = 1\{XU_1 + Z_1\theta_1 + Z_2\theta_2 \geq U_2\}, \tag{M2}$$

$$Y = 1\{X\theta_1 + Z_1\theta_2 + Z_2\theta_3 \geq U\}. \tag{M3}$$

Recall that the extension of our procedure to cover model (M1) was discussed briefly at the end of Section 3.1. Indeed, under model (M1) the index function φ need not even be explicitly specified and it may not satisfy the linearity assumption made under Assumption (2.1). This makes model (M1) the most flexible. Models (M2) and (M3) impose linearity of φ in the latent variables and in the parameters. Here we distinguish two cases. In the first case, (M2) regards (U_1, U_2) as the latent variables in the model. Model (M3) is the same as (M2) except that we have replaced the random slope coefficient U_1 from (M2) with a fixed coefficient. Model (M3) represents the additively separable linear index model that is commonly used in the empirical literature, except for the fact that we do not assume a parametric distribution for U and do not have a model for how the endogenous variable X is generated.

In the presence of an independent and identically distributed random sample, our method can be employed using a slightly modified version of simple plug-in estimators for all probabilities depending on the observed random variables Y, X, Z_1, Z_2 and Z_3. In Appendix B.3 we present a consistency result specifically designed for plug-in estimation in the kinds of problems considered in this paper.[27] However, it is well-known that simple plug-in estimators for the objective function and constraints defining the linear programs in Theorem 3.2 can produce an estimate of the identified set that is inwardly biased (c.f. Chernozhukov et al. (2013)).[28] In addition to the plug-in estimators, we use a half-median unbiased estimator constructed using the inference procedure of Cho and Russell (2020) and report them for comparison. Our confidence sets are also constructed using this procedure. Details on how to use the procedure of Cho and Russell (2020) to construct half-median unbiased estimators and confidence sets in our setting are presented in Appendix B.4, where we discuss how to adapt the inference procedure in Cho and Russell (2020) to accommodate for our profiling procedure. We have also made a few adjustments to the procedure in Cho and Russell (2020) for computational reasons. In particular, the procedure of Cho and Russell (2020) proceeds by bootstrapping linear programs. For specifications (A5) - (A8), a noticeable proportion of these bootstrap linear programs had empty feasible regions, although overall none of the specifications were rejected at the 10% level (i.e. the 90% confidence set for each specification was always non-empty). However, empty feasible regions can slow down the procedure of Cho and Russell (2020), and so for computational reasons we relaxed some of the troublesome constraints when implementing the procedure. Unfortunately our use of a relaxation procedure for some specifications implies that two models with nested assumptions do not necessarily have nested half-median unbiased bounds or confidence sets, as one would expect. It also makes the half-median unbiased estimates and the confidence sets slightly wider than necessary. Despite this, we continue to report the half-median unbiased estimates and confidence sets for comparison with our modified plug-in estimates, although these remarks should be kept in mind when interpreting the results.

[27] Importantly, our consistency result requires a slight (but vanishing) relaxation of the constraint set in our linear programs; in particular, see the sequence "b_n" in Appendix B.3. However, the scale of this sequence can be taken to be extremely small, and so has a minimal impact on the estimated bounds.

[28] This can be proven by a simple application of Jensen's inequality.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	(A8)
(M1): Nonseparability of φ								
Plug-in	$[-0.91, 0.73]$	$[-0.55, 0.73]$	$[-0.81, 0.59]$	$[-0.55, 0.59]$	$[-0.61, 0.46]$	$[-0.38, 0.42]$	$[-0.81, 0.59]$	$[-0.52, 0.58]$
Half-Median	$[-0.91, 0.73]$	$[-0.55, 0.73]$	$[-0.81, 0.59]$	$[-0.55, 0.59]$	$[-0.73, 0.55]$	$[-0.53, 0.57]$	$[-0.85, 0.62]$	$[-0.56, 0.62]$
90% c.s.	$[-0.92, 0.74]$	$[-0.56, 0.74]$	$[-0.82, 0.60]$	$[-0.56, 0.60]$	$[-0.75, 0.57]$	$[-0.54, 0.59]$	$[-0.85, 0.63]$	$[-0.57, 0.64]$
(M2): Linearity of φ (with random coefficients)								
Plug-in	$[-0.91, 0.73]$	$[-0.53, 0.73]$	$[-0.64, 0.40]$	$[-0.37, 0.40]$	$[-0.37, 0.30]$	$[0.09, 0.29]$	$[-0.59, 0.41]$	$[-0.35, 0.41]$
Half-Median	$[-0.98, 0.88]$	$[-0.78, 0.88]$	$[-0.65, 0.43]$	$[-0.73, 0.45]$	$[-0.59, 0.40]$	$[-0.35, 0.39]$	$[-0.66, 0.46]$	$[-0.41, 0.46]$
90% c.s.	$[-0.98, 0.88]$	$[-0.78, 0.88]$	$[-0.66, 0.44]$	$[-0.73, 0.46]$	$[-0.63, 0.42]$	$[-0.37, 0.41]$	$[-0.68, 0.47]$	$[-0.42, 0.47]$
(M3): Linearity of φ (with fixed coefficients)								
Plug-in	$[-0.91, 0.73]$	$[-0.53, 0.73]$	$[-0.64, 0.40]$	$[0.00, 0.40]$	$[0.09, 0.28]$	$[0.09, 0.28]$	$[0.02, 0.41]$	$[0.02, 0.41]$
Half-Median	$[-0.91, 0.73]$	$[-0.53, 0.73]$	$[-0.64, 0.40]$	$[0.00, 0.40]$	$[0.00, 0.41]$	$[0.00, 0.40]$	$[-0.01, 0.45]$	$[-0.01, 0.45]$
90% c.s.	$[-0.92, 0.74]$	$[-0.54, 0.74]$	$[-0.65, 0.41]$	$[0.00, 0.41]$	$[-0.01, 0.43]$	$[-0.02, 0.42]$	$[-0.02, 0.47]$	$[-0.02, 0.47]$

Table 1: Identified sets for the average treatment effect under different specifications and under various assumptions. For the plug-in estimates, we convert all equality constraints to two inequality constraints and introduce a small slackness $b_n = 0.0001/\sqrt{\log(n)}$ which is needed for consistency (see Appendix B.3). Half-median unbiased estimates and a 90% confidence set are also reported. These sets are computed using 999 bootstrap samples using the inference approach in Cho and Russell (2020).

The identified set for μ_{ate} under assumptions (A1) - (A8) and models (M1) - (M3) are reported in Table 1. For simplicity, we report the convex hull of the estimated identified set for each specification. Table 1 also reports our modified plug-in estimator (see Appendix B.3) as well as half-median unbiased estimators and 90% confidence sets. Due to a confluence of factors—including the dimension of the empirical choice probability vector, the large number of constraints, and the sample size—we find that the bootstrap standard errors from our modification of the Cho and Russell (2020) procedure are small, resulting in half-median unbiased estimates that are only slightly more narrow than the 90% confidence sets. Unsurprisingly, the plug-in bounds on μ_{ate} shrink as the strength of our assumptions increase. The most flexible model is (M1) under assumption (A1). It is interesting to note that the bounds on μ_{ate} in this case are contained strictly within the interval $[-1, 1]$, implying that even with the most flexible model, the data provide some information about the average treatment effects. Also note that the identified set for μ_{ate} always overlaps zero for model (M1). As expected, independence of the instrument Z_3 is a stronger assumption than the relaxed independence, hence the plug-in estimates of the identified set under assumptions (A7) and (A8) always contain the plug-in estimates of the identified set under (A5) and (A6). The results also show that relaxed independence does not provide much identifying power (compare the results under Assumptions (A3) and (A7)). On the other hand, independence of Z_3 does induce a noticeable narrowing of the identified set for μ_{ate} (compare the results under Assumptions (A3) and (A6)). The results for this model are a useful benchmark to compare with cases where we impose linearity on the index function.

Next, we see in Table 1 that the linear models from (M2) and (M3) narrow the bounds relative to the case of general nonseparability. Unsurprisingly, the smallest interval for μ_{ate} is obtained under Assumptions (A5) and (A6) for model (M3). For models (M2) and (M3) we make use of our method for profiling θ, as described in Section 3.3. In model (M2) we must profile on $\theta \in \mathbb{R}^2$ and there are 8 representative points. Interestingly, we find that under Assumptions (A1) - (A4) and (A7) - (A8), the identified set of θ is the entire euclidean space $\mathbb{R}^2$. This illustrates that non-trivial bounds on μ_{ate} are possible even when the structural parameters are unidentified. Figure 2 shows the intervals computed using the linear programs of the form (3.13) and (3.14) for each representative point of θ under our various assumptions. The results in Table 1 for model (M2) represent the (convex hull of the) union of the intervals in Figure 2.

In the second linear model (M3), all coefficients are fixed. Thus, we need to profile on a parameter vector $\theta \in \mathbb{R}^3$. Our profiling procedure from Section 3.3 returns 96 representative points, each associated with a polyhedral cone in $\mathbb{R}^3$. Under Assumptions (A1) and (A2), the identified set for θ is $\mathbb{R}^3$, while for all other assumptions (A3) - (A8) we get an informative identified set for θ. In Figure 3 we also show the intervals computed using the linear programs of the form (3.13) and (3.14) for each representative point of θ under our various assumptions. The results in Table 1 for model (M3) represent the (convex hull of the) union of the intervals in Figure 3.

A few interesting patterns emerge when we consider parameters other than the average treatment effect. In particular, consider the counterfactual choice probability:

$$\mu_{ccp}(y) := \sum_{z\in\mathcal{Z}} P_{U|Y,X,Z}(\varphi(1,z,u,\theta) \geq 0 \mid Y = y, X = 0, Z = z) P(Z = z \mid Y = y, X = 0),$$

for $y \in \{0, 1\}$. We focus on the parameter $\mu_{ccp}(0)$ for simplicity, which represents the counterfactual choice probability of visiting a doctor when given private health insurance for the set of individuals who have no insurance and who have chosen not to visit a doctor, averaged across health and marital status. Table 2 reports the convex hull of the estimated identified set for $\mu_{ccp}(0)$ under various model specifications and under various assumptions. Similar to the bounds for μ_{ate}, the half-median unbiased estimates are only slightly more narrow than the 90% confidence sets. We also see that the bounds on counterfactual choice probabilities tend to be wide and uninformative for most assumptions. Note that under Assumption (A1) we always obtain the interval $[0, 1]$ for the estimated identified set. The narrowest bounds are found in model (M3) under Assumptions (A5) and (A6). These bounds allow us to conclude that the probability an individual visits a doctor when given private health insurance, given that they have no private health insurance and did not visit a doctor, is somewhere in the interval $[0.04, 0.19]$.

To summarize, Table 1 shows that most specifications do not identify the sign of μ_{ate}, and Table 2 shows that most bounds on counterfactual choice probabilities are not informative. Exceptions typically occur only under the strongest independence assumptions, given by assumptions (A5) and (A6), and the strongest functional form assumptions, given in model (M3). However, even the strongest set of assumptions considered here are much weaker than the typical assumptions employed in empirical work. For the sake of comparison to our results, we estimate the following bivariate probit model:

$$Y = 1\{X\theta_1 + Z_1\theta_2 + Z_2\theta_3 \geq \varepsilon_1\},$$
$$X = 1\{Z_1\gamma_1 + Z_2\gamma_2 + Z_3\gamma_3 \geq \varepsilon_2\},$$

where (Z_1, Z_2, Z_3) are assumed to be independent from $(\varepsilon_1, \varepsilon_2)$, which are bivariate normal with mean zero,

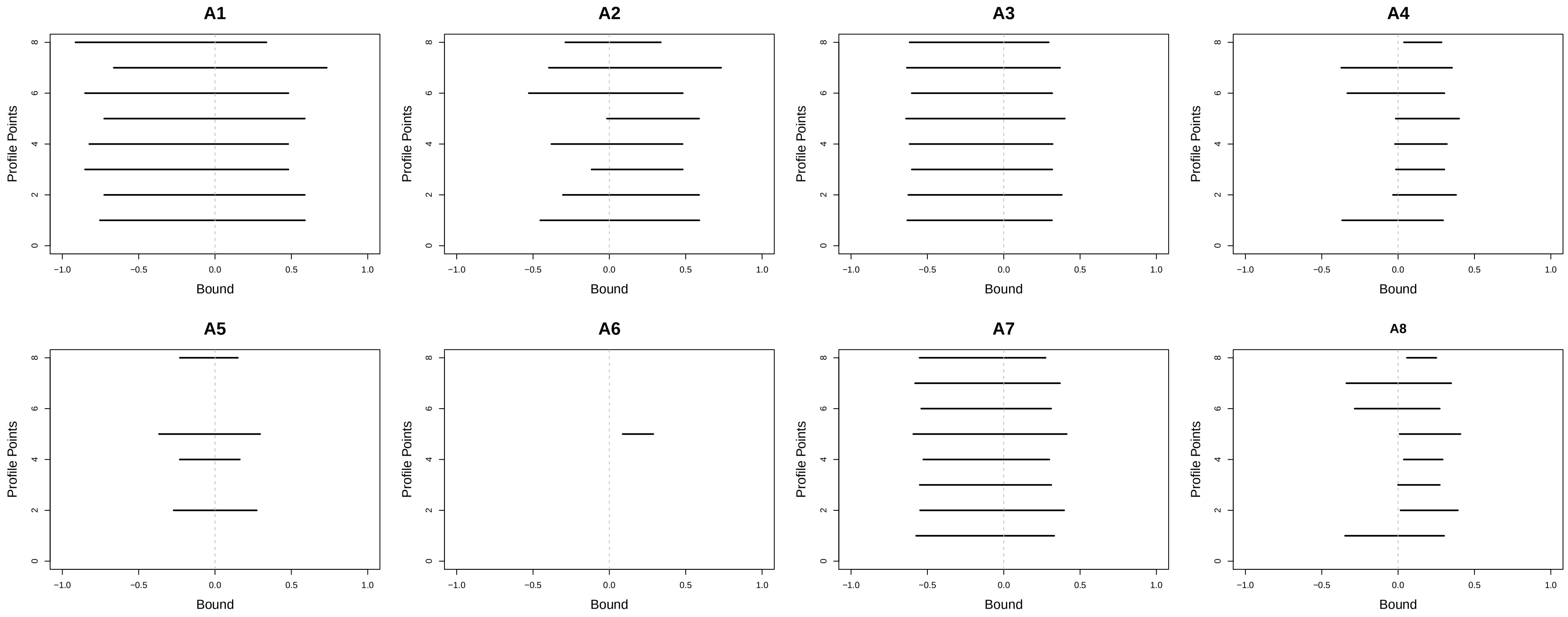


Figure 2: This figure shows the intervals computed using the linear programs of the form (3.13) and (3.14) for each representative point of $\theta \in \mathbb{R}^2$ when bounding μ_{ate} for Model (M2) under various assumptions. The active assumptions are given at the top of each illustration. The axes labelled “profile” correspond to various representative points.

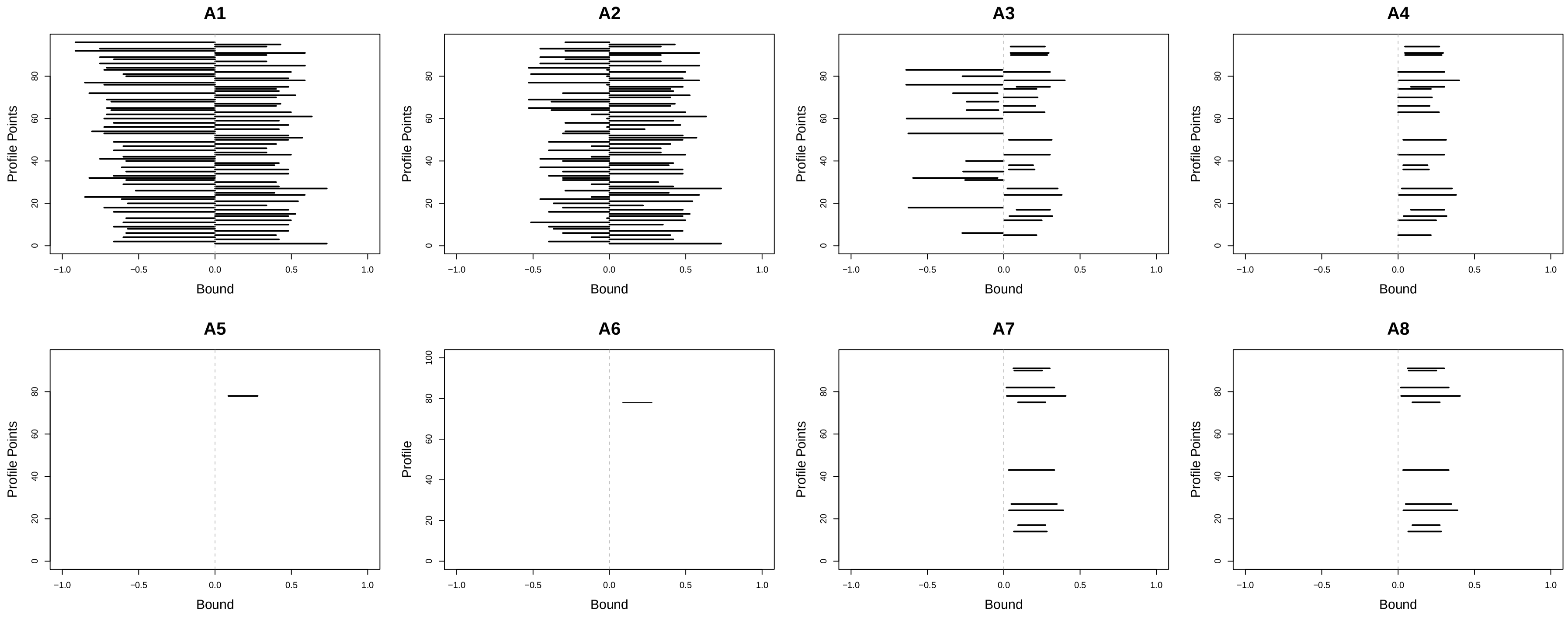


Figure 3: This figure shows the intervals computed using the linear programs of the form (3.13) and (3.14) for each representative point of $\theta \in \mathbb{R}^2$ when bounding μ_{ate} for Model (M3) under various assumptions. The active assumptions are given at the top of each illustration. The axes labelled "profile" correspond to various representative points.

	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)	(A7)	(A8)
(M1): Nonseparability of φ								
Plug-in	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.98]$	$[0.00, 0.98]$	$[0.03, 0.79]$	$[0.03, 0.71]$	$[0.02, 0.95]$	$[0.02, 0.92]$
Half-Median	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.98]$	$[0.00, 0.98]$	$[0.00, 0.93]$	$[0.00, 0.95]$	$[0.00, 0.99]$	$[0.00, 0.97]$
90% c.s.	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 1.000]$	$[0.00, 0.96]$	$[0.00, 0.97]$	$[0.00, 1.00]$	$[0.00, 0.97]$
(M2): Linearity of φ (with random coefficients)								
Plug-in	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.34]$	$[0.00, 0.31]$	$[0.04, 0.23]$	$[0.04, 0.20]$	$[0.02, 0.40]$	$[0.02, 0.36]$
Half-Median	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.46]$	$[0.00, 0.34]$	$[0.00, 0.44]$	$[0.00, 0.40]$	$[0.00, 0.52]$	$[0.00, 0.46]$
90% c.s.	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.47]$	$[0.00, 0.36]$	$[0.00, 0.48]$	$[0.00, 0.43]$	$[0.00, 0.54]$	$[0.00, 0.50]$
(M3): Linearity of φ (with fixed coefficients)								
Plug-in	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.31]$	$[0.00, 0.31]$	$[0.04, 0.19]$	$[0.04, 0.19]$	$[0.02, 0.36]$	$[0.02, 0.36]$
Half-Median	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.31]$	$[0.00, 0.31]$	$[0.00, 0.33]$	$[0.00, 0.31]$	$[0.00, 0.45]$	$[0.00, 0.45]$
90% c.s.	$[0.00, 1.00]$	$[0.00, 1.00]$	$[0.00, 0.32]$	$[0.00, 0.32]$	$[0.00, 0.37]$	$[0.00, 0.35]$	$[0.00, 0.50]$	$[0.00, 0.50]$

Table 2: This table reports the convex hull of the estimated bounds on $\mu_{ccp}(0)$, the counterfactual choice probability of visiting doctor when granted insurance for those who chose not to visit a doctor without insurance. For the plug-in estimates, we convert all equality constraints into two inequality constraints and introduce a small slackness $b_n = 0.0001/\sqrt{\log(n)}$, which is needed for consistency (see Appendix B.3). Half-median unbiased estimates and a 90% confidence set are also reported. These sets are computed using 999 bootstrap samples using the inference approach in Cho and Russell (2020).

unit variance and correlation ρ. This model was estimated with our data using maximum likelihood, and μ_{ate} was estimated as 0.16 with a bootstrapped confidence interval of $[0.11, 0.20]$. This value for μ_{ate} lies within all of our bounds in Table 1, and seems to suggest strong evidence of a positive causal effect of health insurance on the decision to visit the doctor.[29] However, the bivariate probit model is highly parameterized, and the results from Table 1 suggest that under weaker assumptions the sign of μ_{ate} may not be identified.

The previous literature studying the effects of health insurance on the utilization of health care services is full of mixed results, and Table 1 suggests that highly parameterized models may give highly significant, but possibly misleading results relative to models that make weaker assumptions.

6 Conclusion

This paper considers (partial) identification of a variety of parameters in binary response models with possibly endogenous regressors. Importantly, our class of models allows for general nonseparability of the index function in latent variables, and does not require any parametric distributional assumptions. Our approach to bounding counterfactual parameters is based on framing the bounding in terms of optimization problems. Our specific partition of the latent variable space is key to our suggested procedure, and we show how to enumerate the sets in this partition using results from the literature on computational geometry and hyperplane arrangements. In doing so, we provide a feasible method of constructing bounds on counterfactual quantities under weak assumptions where the latent variables may be multi-dimensional and nonseparable.

[29] Han and Lee (2019) obtain a similar result in a model allowing for ε_1 and ε_2 to have unrestricted marginals, and a flexible dependence structure. However, they consider a different model from us, and the average treatment effect in Han and Lee (2019) is different from ours; we consider the average treatment effect averaged over all values of (x, z), while they report the average treatment effect at the average value of their conditioning variables. They report the average treatment effect at various quantiles of their conditioning variables.

We also thoroughly study the special case when the index function is linear in parameters, and show how to compute exact (i.e. not approximate) sharp bounds on counterfactual quantities, as well as how to adapt a recent inference procedure to the setting in this paper in order to construct confidence sets and bias-corrected estimates of the identified set. Finally, we show how to impose independence and monotonicity assumptions, and we present an application of our method to study the effects of private health insurance on the utilization of health care services.

There are a number of obvious further directions in which to expand the ideas presented in this paper. For example, the consideration of multinomial choice models, triangular systems, or general simultaneous discrete choice models all seem to be natural next steps. In addition, a major emphasis in this paper, as in other recent papers, is on the important computational problems that arise in models that are partially identified. We believe exploring applications of state-of-the-art algorithms in computer science to problems in econometrics—as we have attempted here—is a fruitful avenue of research.

References

Aliprantis, C. D. and Border, K. C. (2006). *Infinite dimensional analysis: a hitchhiker's guide*. Springer.

Allen, R. and Rehbeck, J. (2019). Identification with additively separable heterogeneity. *Econometrica*, 87(3):1021–1054.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Artstein, Z. (1983). Distributions of random sets and random selections. *Israel Journal of Mathematics*, 46(4):313–324.

Avis, D., Bremner, D., and Seidel, R. (1997). How good are convex hull algorithms? *Computational Geometry*, 7(5-6):265–301.

Avis, D. and Fukuda, K. (1996). Reverse search for enumeration. *Discrete applied mathematics*, 65(1-3):21–46.

Bajari, P., Dalton, C., Hong, H., and Khwaja, A. (2014). Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *The RAND Journal of Economics*, 45(4):747–763.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier.

Bennett, J. F. (1956). Determination of the number of independent parameters of a score matrix from the examination of rank orders. *Psychometrika*, 21(4):383–393.

Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79(6):1785–1821.

Beresteanu, A., Molchanov, I., and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, 166(1):17–32.

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *Review of Economic Studies*, 71(3):655–679.

Blundell, R. W. and Smith, R. J. (1989). Estimation in a class of simultaneous equation limited dependent variable models. *The Review of Economic Studies*, 56(1):37–57.

Bremner, D. (1999). Incremental convex hull algorithms are not output sensitive. *Discrete & Computational Geometry*, 21(1):57–68.

Buck, R. (1943). Partition of space. *The American Mathematical Monthly*, 50:541–544.

Cardon, J. H. and Hendel, I. (2001). Asymmetric information in health insurance: evidence from the national medical expenditure survey. *RAND Journal of Economics*, pages 408–427.

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

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013). Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737.

Chesher, A. (2013). Semiparametric structural models of binary response: shape restrictions and partial identification. *Econometric Theory*, pages 231–266.

Chesher, A. and Rosen, A. M. (2014). An instrumental variable random-coefficients model for binary outcomes. *The econometrics journal*, 17(2):S1–S19.

Chesher, A. and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3):959–989.

Chesher, A. and Rosen, A. M. (2019). Generalized instrumental variable models methods and applications. Technical report, cemmap working paper.

Chesher, A., Rosen, A. M., and Smolinski, K. (2013). An instrumental variable model of multiple discrete choice. *Quantitative Economics*, 4(2):157–196.

Chiburis, R. C. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, 159(2):267–275.

Chiong, K., Hsieh, Y.-W., and Shum, M. (2017). Counterfactual estimation in semiparametric discrete-choice models. *Available at SSRN 2979446.*

Cho, J. and Russell, T. M. (2020). Simple inference on functionals of set-identified parameters defined by linear moments. *arXiv preprint arXiv:1810.03180.*

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334.

Cover, T. M. (1967). The number of linearly inducible orderings of points in d-space. *SIAM Journal on Applied Mathematics*, 15(2):434–439.

Dong, Y. and Lewbel, A. (2015). A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews*, 34(1-2):82–105.

Durrett, R. (2010). *Probability: theory and examples, fourth edition.* Cambridge university press.

Finkelstein, A. and McGarry, K. (2006). Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review*, 96(4):938–958.

Fukuda, K. (2014). *Frequently asked questions in polyhedral computation: https://people.inf.ethz.ch/fukudak//polyfaq/polyfaq.html.*

Fukuda, K. and Prodon, A. (1995). Double description method revisited. In *Franco-Japanese and Franco-Chinese Conference on Combinatorics and Computer Science*, pages 91–111. Springer.

Galichon, A. (2016). *Optimal transport methods in economics.* Princeton University Press.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Geyer, C. (2019). *Using the RCDD package: https://cran.r-project.org/web/packages/rcdd/vignettes/vinny.pdf.*

Gu, J. and Koenker, R. (2020). Nonparametric maximum likelihood methods for binary response models with random coefficients. *Journal of the American Statistical Association*, pages 1–47.

Gu, J. and Russell, T. M. (2021). Partial identification in nonseparable binary response models with endogenous regressors. *arXiv preprint arXiv:2101.01254.*

Gunsilius, F. F. (2020). A path-sampling method to partially identify causal effects in instrumental variable models. Working paper.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Han, S. and Lee, S. (2019). Estimation in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Applied Econometrics*, 34(6):994–1015.

Heckman, J. J. and Pinto, R. (2018). Unordered monotonicity. *Econometrica*, 86(1):1–35.

Himmelberg, C. (1975). Measurable relations, fundam.

Hirano, K. and Porter, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790.

Ichimura, H. and Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Kohler, D. A. (1967). Projections of convex polyhedral sets. Technical report, University of California Berkeley Operations Research Center.

Lafférs, L. (2019). Identification in models with discrete variables. *Computational Economics*, 53(2):657–696.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177.

Lewbel, A., Dong, Y., and Yang, T. T. (2012). Comparing features of convenient estimators for binary choice models with endogenous regressors. *Canadian Journal of Economics/Revue canadienne d'économique*, 45(3):809–829.

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

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review*, 48(4):1393–1410.

Manski, C. F. and Pepper, J. V. (1998). Monotone instrumental variables with an application to the returns to schooling. Technical report, National Bureau of Economic Research.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica: Journal of the Econometric Society*, pages 239–270.

Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica*, 71(5):1339–1375.

Molchanov, I. (2017). *Theory of random sets*. Springer Science & Business Media.

Molchanov, I. S. (1998). A limit theorem for solutions of inequalities. *Scandinavian Journal of Statistics*, 25(1):235–242.

Motzkin, T., Raiffa, H., Thompson, G., and Thrall, R. (1953). The double description method. In Kuhn, H. and Tucker, A., editors, *Contributions to theory of games*. Princeton University Press.

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

Norberg, T. (1992). On the existence of ordered couplings of random sets—with applications. *Israel Journal of Mathematics*, 77(3):241–264.

Pearl, J. (2009). *Causality*. Cambridge university press.

Rada, M. and Cerny, M. (2018). A new algorithm for enumeration of cells of hyperplane arrangements and a comparison with avis and fukuda's reverse search. *SIAM Journal on Discrete Mathematics*, 32(1):455–473.

Rothschild, M. and Stiglitz, J. (1978). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. In *Uncertainty in economics*, pages 257–280. Elsevier.

Russell, T. M. (2019). Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, pages 1–15.

Sainte-Beuve, M.-F. (1974). On the extension of von neumann-aumann's theorem. *Journal of Functional Analysis*, 17(1):112–129.

Shaikh, A. M. and Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79(3):949–955.

Sleumer, N. H. (1999). Output-sensitive cell enumeration in hyperplane arrangements. *Nordic journal of computing*, 6(2):137–147.

Tebaldi, P., Torgovitsky, A., and Yang, H. (2019). Nonparametric estimates of demand in the california health insurance exchange. Technical report, National Bureau of Economic Research.

Torgovitsky, A. (2019). Partial identification by extending subdistributions. *Quantitative Economics*, 10(1):105–144.

Vytlacil, E. and Yildiz, N. (2007). Dummy endogenous variables in weakly separable models. *Econometrica*, 75(3):757–779.

A Proofs

A.1 Proofs of Results in the Main Text

Proof of Theorem 2.1. Let $\mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$ denote the set of all conditional distributions $P_{Y_\gamma|Y,X,Z}$ such that there exists a pair $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying:

$$P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z), U, \theta) \geq 0 \mid Y = y, X = x, Z = z\right),$$

$P_{X,Z}-$a.s. To prove the result it suffices to show $\mathcal{P}^*_{Y_\gamma|Y,X,Z} = \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$. To do this, we show that $\mathcal{P}^*_{Y_\gamma|Y,X,Z} \subset \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$ and $\mathcal{P}^{**}_{Y_\gamma|Y,X,Z} \subset \mathcal{P}^*_{Y_\gamma|Y,X,Z}$. To this end, begin by fixing an arbitrary $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$. By Definition 2.2 we have:

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z), U, \theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = 1, \tag{A.1}$$

$P_{Y,X,Z,U}-$a.s. for some $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$. For this pair $(P_{U|Y,X,Z}, \theta)$ we have:

$$\begin{aligned}
&P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z, U = u\right) \\
&= P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = 1, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z), U, \theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right),
\end{aligned}$$

$P_{Y,X,Z,U}-$a.s., which follows from (A.1). Now note:

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = 1, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z), U, \theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = \mathbb{1}\{\varphi(\gamma(x,z), U, \theta) \geq 0\},$$

$P_{Y,X,Z,U}-$a.s. Thus we have:

$$\begin{aligned}
P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) &= \int P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = 1 \mid Y = y, = x, Z = z, U = u\right) dP_{U|Y,X,Z} \\
&= \int \mathbb{1}\{\varphi(\gamma(x,z), U, \theta) \geq 0\} \, dP_{U|Y,X,Z} \\
&= P_{U|Y,X,Z}(\varphi(\gamma(X,Z), U, \theta) \geq 0 \mid Y = y, X = x, Z = z),
\end{aligned}$$

$P_{Y,X,Z}-$a.s. In other words, for our $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$ we have shown that there exists a pair $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying:

$$P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z), U, \theta) \geq 0 \mid Y = y, X = x, Z = z\right),$$

$P_{Y,X,Z}-$a.s. This proves $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$, and since $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$ was arbitrary we conclude that $\mathcal{P}^*_{Y_\gamma|Y,X,Z} \subset \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$.

For the reverse inclusion, fix any arbitrary $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$. Then by definition there exists a pair $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying:

$$P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z), U, \theta) \geq 0 \mid Y = y, X = x, Z = z\right),$$

$P_{Y,X,Z}-$a.s. It suffices to show that for this pair $(P_{U|Y,X,Z}, \theta)$ there exists $P_{Y_\gamma|Y,X,Z,U}$ satisfying:

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = 1, \tag{A.2}$$

$P_{Y,X,Z,U}-$a.s. By the Radon-Nikodym Theorem, the existence of a (version of) $P_{Y_\gamma|Y,X,Z,U}$ is guaranteed by the fact that $P_{Y_\gamma,U|Y,X,Z} \ll P_{U|Y,X,Z}$. Since all spaces involved are euclidean, we can choose the version to be an almost surely unique regular conditional distribution (c.f. Durrett (2010) Theorem 5.1.9). By construction this $P_{Y_\gamma|Y,X,Z,U}$ satisfies:

$$\begin{aligned} &P_{Y_\gamma,U|Y,X,Z}(Y_\gamma \in A, U \in B \mid Y = y, X = x, Z = z) \\ &\qquad = \int_B P_{Y_\gamma|Y,X,Z,U}(Y_\gamma \in A \mid Y = y, Z = z, X = x, U = u)\, dP_{U|Y,X,Z}, \end{aligned}$$

$P_{Y,X,Z}-$a.s. for every $A \subset \{0,1\}$ and $B \in \mathfrak{B}(\mathcal{U})$. Now note that:

$$\begin{aligned} P_{Y_\gamma|Y,X,Z,U}(Y_\gamma = 1, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u) = \mathbb{1}\{\varphi(\gamma(x,z),u,\theta) \geq 0\}, \\ P_{Y_\gamma|Y,X,Z,U}(Y_\gamma = 0, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u) = \mathbb{1}\{\varphi(\gamma(x,z),u,\theta) < 0\}. \end{aligned}$$

$P_{Y,X,Z}-$a.s. Thus:

$$\begin{aligned} &P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z\right) \\ &= \int_{\mathcal{U}} P_{Y_\gamma|Y,X,Z,U}(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u)\, dP_{U|Y,X,Z} \\ &= \int_{\mathcal{U}} P_{Y_\gamma|Y,X,Z,U}(Y_\gamma = 1, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u)\, dP_{U|Y,X,Z} \\ &\qquad + \int_{\mathcal{U}} P_{Y_\gamma|Y,X,Z,U}(Y_\gamma = 0, Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u)\, dP_{U|Y,X,Z} \\ &= \int_{\mathcal{U}} \mathbb{1}\{\varphi(\gamma(x,z),u,\theta) \geq 0\}\, dP_{U|Y,X,Z} + \int_{\mathcal{U}} \mathbb{1}\{\varphi(\gamma(x,z),u,\theta) < 0\}\, dP_{U|Y,X,Z} \\ &= P_{U|Y,X,Z}(\varphi(\gamma(x,z),u,\theta) \geq 0 \mid Y = y, X = x, Z = z) + P_{U|Y,X,Z}(\varphi(\gamma(x,z),u,\theta) < 0 \mid Y = y, X = x, Z = z) \\ &= 1, \end{aligned}$$

$P_{Y,X,Z}-$a.s. This proves (A.2) and thus shows $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$. Since $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$ was arbitrary we can conclude that $\mathcal{P}^{**}_{Y_\gamma|Y,X,Z} \subset \mathcal{P}^*_{Y_\gamma|Y,X,Z}$. Combining the two inclusions, we have $\mathcal{P}^*_{Y_\gamma|Y,X,Z} = \mathcal{P}^{**}_{Y_\gamma|Y,X,Z}$. This completes the proof.

■

Proof of Theorem 3.1. Let $P_{Y_\gamma|Y,X,Z}$ be a collection of conditional distributions, and suppose there exists $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying (2.5). Note that (3.7) is equivalent to (2.5), so we can conclude that $(P_{U|Y,X,Z}, \theta)$ satisfies (3.7). Furthermore, by definition $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ implies that:

$$P_{U|Y,X,Z}(U \in \mathcal{U}(Y,X,Z,\theta) \mid Y = y, X = x, Z = z) = 1,\ P_{Y,X,Z} - a.s.,$$

which is equivalent to conditions (3.5) and (3.6). This shows that any pair $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$ satisfying (2.5) satisfies (3.5) - (3.7).

For the reverse, fix any $\theta \in \Theta$ and any collection $P_{U|Y,X,Z}$ of probability measures on the sets in $\mathcal{A}(\theta)$ satisfying (3.5) - (3.7). We show that $P_{U|Y,X,Z}$ can be extended to a (not necessarily unique) probability measure $\tilde{P}_{U|Y,X,Z}$ on $\mathfrak{B}(\mathcal{U})$ in a manner that ensures $\tilde{P}_{U|Y,X,Z}$ satisfies (2.5) and such that $(\tilde{P}_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$. Furthermore, by the definition of an extension, $\tilde{P}_{U|Y,X,Z}$ agrees with $P_{U|Y,X,Z}$ on all sets of the form $\mathcal{A}(\theta)$. To construct the extension, note that the sets in $\mathcal{A}(\theta)$ form a disjoint partition of $\mathcal{U}$. Now select a single point $u(s,\theta)$ from the interior of each set $\mathcal{U}(s,\theta)$ in the collection $\mathcal{A}(\theta)$; if $\mathcal{U}(s,\theta)$ has empty interior, choose $u(s,\theta)$ as an arbitrary point from $\mathcal{U}$. For any set $A \subset \mathcal{U}$, define the indicator:

$$\mathbb{1}(A,\theta,s) = \mathbb{1}\{u(s,\theta) \in A \cap \text{int}(\mathcal{U}(s,\theta))\}.$$

Now define the function $\mu_{y,x,z} : \mathfrak{B}(\mathcal{U}) \to \mathbb{R}$ as:

$$\mu_{y,x,z}(B) := \sum_{s\in\{0,1\}^m} \mathbb{1}(B,\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right).$$

To verify that this is a proper probability measure on $\mathfrak{B}(\mathcal{U})$, we must show that (i) $\mu_{y,x,z}(B) \geq \mu_{y,x,z}(\varnothing) = 0$ for every $B \in \mathfrak{B}(\mathcal{U})$, (ii) $\mu_{y,x,z}(\mathcal{U}) = 1$, and (iii) for any countable sequence of disjoint sets $\{A_i\}_{i=1}^\infty$ in $\mathfrak{B}(\mathcal{U})$, we have:

$$\mu_{y,x,z}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu_{y,x,z}(A_i).$$

The first property holds since $\mathbb{1}(\varnothing,\theta,s) = 0$ for all s. To verify the second property, note that $\mathbb{1}(\mathcal{U},\theta,s) = 1$ for all s, so that:

$$\begin{aligned}
\mu_{y,x,z}(\mathcal{U}) &= \sum_{s\in\{0,1\}^m} \mathbb{1}(\mathcal{U},\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right) \\
&= \sum_{s\in\{0,1\}^m} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right) \\
&= 1,
\end{aligned}$$

where the last line holds since $P_{U|Y,X,Z}$ is a probability measure on $\mathcal{A}(\theta)$. For the third property, note that for two disjoint Borel sets $A_1, A_2 \in \mathfrak{B}(\mathcal{U})$ we have:

$$\mathbb{1}(A_1 \cup A_2,\theta,s) = \mathbb{1}(A_1,\theta,s) + \mathbb{1}(A_2,\theta,s).$$

Inducting on this formula, we conclude that for countable disjoint sets $\{A_i\}_{i=1}^\infty$ in $\mathfrak{B}(\mathcal{U})$, we have:

$$\mathbb{1}\left(\bigcup_{i=1}^\infty A_i,\theta,s\right) = \sum_{i=1}^\infty \mathbb{1}(A_i,\theta,s),$$

Thus we can conclude:

$$\begin{aligned}\mu_{y,x,z}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{s\in\{0,1\}^m} \mathbb{1}\left(\bigcup_{i=1}^{\infty} A_i, \theta, s\right) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right)\\ &= \sum_{s\in\{0,1\}^m}\sum_{i=1}^{\infty} \mathbb{1}(A_i,\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right)\\ &= \sum_{i=1}^{\infty}\sum_{s\in\{0,1\}^m} \mathbb{1}(A_i,\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z\right)\\ &= \sum_{i=1}^{\infty} \mu_{y,x,z}(A_i).\end{aligned}$$

Thus, our measure satisfies countable additivity. We conclude that $\mu_{y,x,z}$ is a proper probability measure. Note that the argument above has been completed for a single triple (y,x,z) indexing the conditioning variables. However, we can repeat the same argument as above for all (y,x,z) assigned positive probability, and thus can construct a corresponding probability measure $\mu_{y,x,z}$ satisfying all the conditions described above for each such (y,x,z).

Now we define $\tilde{P}_{U|Y,X,Z} : \mathfrak{B}(\mathcal{U}) \to [0,1]$ by $\tilde{P}_{U|Y,X,Z}(B \mid Y=y, X=x, Z=z) = \mu_{y,x,z}(B)$ for all $B \in \mathfrak{B}(\mathcal{U})$ and all (y,x,z) assigned positive probability. By the above, $\tilde{P}_{U|Y,X,Z}(\,\cdot \mid Y=y, X=x, Z=z)$ is a proper probability measure on $\mathfrak{B}(\mathcal{U})$ for each (y,x,z). Also note that for any triple $(1,x,z)$ assigned positive probability, the pair $(\tilde{P}_{U|Y,X,Z},\theta)$ satisfies:

$$\begin{aligned}&\tilde{P}_{U|Y,X,Z}(\mathcal{U}(1,x,z,\theta) \mid Y=1, X=x, Z=z)\\ &= \sum_{s\in S_j} \tilde{P}_{U|Y,X,Z}(\mathcal{U}(s,\theta) \mid Y=1, X=x, Z=z)\\ &= \sum_{s\in S_j}\sum_{s'\in\{0,1\}^n} \mathbb{1}(\mathcal{U}(s,\theta),\theta,s') P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=1, X=x, Z=z\right)\\ &= \sum_{s\in S_j} \mathbb{1}(\mathcal{U}(s,\theta),\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=1, X=x, Z=z\right)\\ &= 1,\end{aligned}$$

which follows from (3.5). Furthermore, for any triple $(0,x,z)$ assigned positive probability, the pair $(\tilde{P}_{U|Y,X,Z},\theta)$ also satisfies:

$$\begin{aligned}&\tilde{P}_{U|Y,X,Z}(\mathcal{U}(0,x,z,\theta) \mid Y=0, X=x, Z=z)\\ &= \sum_{s\in S_j^c} \tilde{P}_{U|Y,X,Z}(\mathcal{U}(s,\theta) \mid Y=0, X=x, Z=z)\\ &= \sum_{s\in S_j^c}\sum_{s'\in\{0,1\}^n} \mathbb{1}(\mathcal{U}(s,\theta),\theta,s') P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=0, X=x, Z=z\right)\\ &= \sum_{s\in S_j^c} \mathbb{1}(\mathcal{U}(s,\theta),\theta,s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=0, X=x, Z=z\right)\end{aligned}$$

$$= 1,$$

which follows from (3.6). Conclude that:

$$\tilde{P}_{U|Y,X,Z}(U \in \mathcal{U}(Y,X,Z,\theta) \mid Y = y, X = x, Z = z) = 1, \ \ a.s.$$

This shows that $(\tilde{P}_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$. Finally, setting $C := \{u \in \mathcal{U} : \varphi(\gamma(x,z), u, \theta) \geq 0\}$, it is straightforward to show that:

$$\begin{aligned}\tilde{P}_{U|Y,X,Z}\left(C \mid Y = y, X = x, Z = z\right) &= \sum_{s \in \{0,1\}^m} \mathbb{1}(C, \theta, s) P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x, Z = z\right) \\ &= \sum_{s \in S_{\gamma(j)}} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x_j, Z = z_j\right) \\ &= P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j\right),\end{aligned}$$

for all (y, x_j, z_j) assigned positive probability, which follows from (3.7). This is exactly condition (2.5). Conclude that $(\tilde{P}_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ and that $(\tilde{P}_{U|Y,X,Z}, \theta)$ satisfies (2.5). This completes the proof.

■

Proof of Theorem 3.2. Note that the constraints in (3.9) are equivalent to the constraints in (3.5) and (3.6). Furthermore, the objective function in the optimization problems in Theorem 3.2 enforce (3.7). Thus, using Theorem 3.1, a distribution $\pi(\theta)$ is feasible in the optimization problems from Theorem 3.2 if and only if there exists a collection of Borel conditional probability measures $P_{U|Y,X,Z}$ satisfying (2.5) with $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$. However, by Theorem 2.1, there exists a collection of Borel conditional probability measures $P_{U|Y,X,Z}$ satisfying (2.5) with $(P_{U|Y,X,Z}, \theta) \in \mathcal{I}^*_{Y,X,Z}$ if and only $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$, where $P_{Y_\gamma|Y,X,Z}$ is the (collection of) conditional distribution(s) satisfying (2.5). ■

Proof of Proposition 3.1. This follows immediately from the results of Buck (1943). ■

A.2 Measurability Results

Definition A.1 (Effros-Measurability, Random Set, Selection). *Let $(\Omega, \mathfrak{A}, P)$ be a probability space, let $\mathcal{V}$ be a Polish space, and let $\mathcal{O}_\mathcal{V}$ denote the collection of all open sets on $\mathcal{V}$. A multifunction $\boldsymbol{V} : \Omega \to \mathfrak{F}_\mathcal{V}$ is called Effros-measurable if for every $A \in \mathcal{O}_\mathcal{V}$ we have $\boldsymbol{V}^-(A) := \{\omega \in \Omega : \boldsymbol{V}(\omega) \cap A \neq \varnothing\} \in \mathfrak{A}$. A random element $V : \Omega \to \mathcal{V}$ is called a (measurable) selection of $\boldsymbol{V}$ if $V(\omega) \in \boldsymbol{V}(\omega)$ for $P-$almost all $\omega \in \Omega$.*

Lemma A.1. *Suppose Assumption 2.1 holds. Then for each $\theta \in \Theta$, the map $\mathcal{U}(\,\cdot\,, \theta) : \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \to \mathcal{U}$ is an Effros-measurable multifunction, and thus is a random set on $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$.*

Proof of Lemma A.1. For any fixed $\theta \in \Theta$ and any open set $A \subset \mathcal{U}$. We have:

$$\{(y,x,z) : \mathcal{U}(y,x,z,\theta) \cap A \neq \varnothing\} = G_0(A) \cup G_1(A),$$

where:

$$G_0(A) := \{(0,x,z) : \mathcal{U}(0,x,z,\theta) \cap A \neq \varnothing\},$$
$$G_1(A) := \{(1,x,z) : \mathcal{U}(1,x,z,\theta) \cap A \neq \varnothing\}.$$

Since $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Z})$ is closed under unions, it suffices to show $G_0(A), G_1(A) \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Z})$. In particular, it suffices to show Effros-measurability of the maps:

$$\mathcal{U}(0,x,z,\theta) = \{u \in \mathcal{U} : \varphi(x,z,u,\theta) < 0\},$$
$$\mathcal{U}(1,x,z,\theta) = \{u \in \mathcal{U} : \varphi(x,z,u,\theta) \geq 0\}.$$

Effros measurability of $\mathcal{U}(0,x,z,\theta)$ follows directly from Lemma 18.7 in Aliprantis and Border (2006) after noting that $\varphi(\cdot,\theta)$ is a Caratheodory function, and $(-\infty,0)$ is an open set. For measurability of $\mathcal{U}(1,x,z,\theta)$, consider an arbitrary point $u_0 \in \mathcal{U}$, and define:

$$d(u_0, \mathcal{U}(1,x,z,\theta)) := \inf_{u \in \mathcal{U}(1,x,z,\theta)} ||u_0 - u||.$$

By Assumption 2.1, the set $\mathcal{U}(1,x,z,\theta)$ is a closed halfspace in $\mathbb{R}^{d_u}$ of the form:

$$\mathcal{U}(1,x,z,\theta) = \left\{u \in \mathcal{U} : -\tilde{\varphi}_1(x,z,\theta)^\top u - \tilde{\varphi}_2(x,z,\theta) \leq 0\right\},$$

for some measurable functions $\tilde{\varphi}_1(x,z,\theta)$ and $\tilde{\varphi}_2(x,z,\theta)$. It follows that:[30]

$$d(u_0, \mathcal{U}(1,x,z,\theta)) = \frac{|-\tilde{\varphi}_1(x,z,\theta)^\top u_0 - \tilde{\varphi}_2(x,z,\theta)|_+}{||\tilde{\varphi}_1(x,z,\theta)||}.$$

so $d(u_0, \mathcal{U}(1,x,z,\theta))$ is itself measurable in (x,z). Following Himmelberg (1975) (see also Theorem 1.3.3 in Molchanov (2017)) this implies that $\mathcal{U}(1,\cdot,\theta) : \mathcal{X} \times \mathcal{Z} \to \mathcal{U}$ is an Effros-measurable multifunction. This completes the proof. ■

Given a $\sigma-$algebra $\mathfrak{F}$ on a space $\mathcal{R}$, the P-completion of $\mathfrak{F}$ is the smallest $\sigma-$algebra containing $\mathfrak{F}$ as well as all $P-$null sets of $\mathcal{R}$. The intersection of all $P-$completions of $\mathfrak{F}$ (over all P) is called the *universal* $\sigma-$*algebra*, and functions that are measurable with respect to the universal $\sigma-$algebra are said to be *universally measurable*. The following Lemma shows that the random set $\mathcal{U}(Y,X,Z,\theta)$ admits a universally measurable selection under Assumption 2.1.

Lemma A.2. *Suppose Assumption 2.1 holds. Then the random set $\mathcal{U}(Y,X,Z,\theta)$ admits a universally measurable selection for every $\theta \in \Theta$ ensuring it is non-empty almost surely.*

Proof of Lemma A.2. Fix some $\theta \in \Theta$ ensuring $\mathcal{U}(Y,X,Z,\theta)$ is almost surely non-empty. By Lemma A.1,

[30] Note this follows from the fact that the distance between a point $\boldsymbol{x}_0$ and the halfspace $H := \{\boldsymbol{x} : \boldsymbol{a}^\top \boldsymbol{x} + b \leq 0\}$ is given by:

$$d(\boldsymbol{x}_0, H) := \frac{|\boldsymbol{a}^\top \boldsymbol{x}_0 + b|_+}{||\boldsymbol{a}||}.$$

$\mathcal{U}(Y,X,Z,\theta)$ is an Effros-measurable multifunction, and by Theorem 1.3.3 in Molchanov (2017) this implies that the graph of $\mathcal{U}(Y,X,Z,\theta)$ belongs to $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Z}) \times \mathfrak{B}(\mathcal{U})$; that is, $\mathcal{U}(Y,X,Z,\theta)$ is graph-measurable. The result then follows immediately from Theorem 3 of Sainte-Beuve (1974). ■

B Additional Definitions and Results

B.1 Independence Assumptions

Under Assumption 4.1, we have the following definition of the identified set, which is analogous to both Definitions 2.1 and 2.2.

Definition B.1. *Under Assumptions 2.1 and 4.1, the identified set $\mathcal{I}^*_{Y,X,Z}$ is the set of all pairs $(P_{U|Y,X,Z},\theta)$ such that:*

(i) $(P_{U|Y,X,Z},\theta)$ satisfies:

$$P_{U|Y,X,Z}(U \in \mathcal{U}(Y,X,Z,\theta) \mid Y=y, X=x, Z=z) = 1,$$

$P_{Y,X,Z}-$a.s.; and

(ii) For all Borel sets $A \in \mathfrak{B}(\mathcal{U})$ we have $P_{U|Z}(A \mid Z=z) = P_U(A)$, P_Z-a.s.

Furthermore, under Assumptions 2.1, 2.2 and 4.1, the identified set of counterfactual conditional distributions $\mathcal{P}^*_{Y_\gamma|Y,X,Z,U}$ is the set of all conditional distributions $P_{Y_\gamma|Y,X,Z,U}$ satisfying:

$$P_{Y_\gamma|Y,X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y=y, X=x, Z=z, U=u\right) = 1,$$

$P_{Y,X,Z,U}-$a.s. for some pair $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$.

Here we do not consider the case when Assumptions 2.1 and 4.2 hold, but we again note that this definition (and the results to follow) are easily modified to accommodate the case when any combination of these assumptions hold. We now provide the following Corollary whose proof follows almost identically to that of Theorems 2.1 and 3.1, with the exception being that we require condition (ii) of Definition B.1 to hold.

Corollary B.1. *Under Assumptions 2.1, 2.2 and 4.1, a counterfactual conditional distribution $P_{Y_\gamma|Y,X,Z}$ satisfies $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$ if and only if there exists a pair $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$ (for $\mathcal{I}^*_{Y,X,Z}$ from Definition B.1) satisfying:*

$$P_{Y_\gamma|X,Z}\left(Y_\gamma = 1 \mid Y=y, X=x, Z=z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z),U,\theta) \geq 0 \mid Y=y, X=x, Z=z\right), \quad \text{(B.1)}$$

$P_{Y,X,Z}-$a.s. Furthermore, for any collection of counterfactual conditional distributions $P_{Y_\gamma|Y,X,Z}$, there exists a collection of Borel conditional probability measures $P_{U|Y,X,Z}$ satisfying (B.1) with $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$

(for $\mathcal{I}^*_{Y,X,Z}$ from Definition B.1) if and only if there exists a collection $P_{U|Y,X,Z}$ of probability measures on the sets in $\mathcal{A}(\theta)$ from (3.4) satisfying:

$$\sum_{s\in S_j} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=1, X=x_j, Z=z_j\right)=1,$$

$$\sum_{s\in S_j^c} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=0, X=x_j, Z=z_j\right)=1,$$

$$\sum_{s\in S_{\gamma(j)}} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x_j, Z=z_j\right)=P_{Y_\gamma|Y,X,Z}\left(Y_\gamma=1 \mid Y=y, X=x_j, Z=z_j\right),$$

for $y\in\{0,1\}$ and $j\in\{1,\dots,m\}$ assigned positive probability, and:

$$\sum_y\sum_x P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z_k\right)P(Y=y,X=x\mid Z=z_k)$$
$$=\sum_y\sum_x P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y=y, X=x, Z=z_{k+1}\right)P(Y=y,X=x\mid Z=z_{k+1}), \quad \text{(B.2)}$$

for all $s\in\{0,1\}^m$ and all $k=1,\dots,m_z-1$ assigned positive probability.

Proof of Corollary B.1. The first statement follows a proof identical to the proof of Theorem 2.1. For the second statement, the forward direction is identical to the proof of Theorem 3.1. The reverse direction is similar to the proof of Theorem 3.1, with the exception that we must show that the extended measure on $\mathfrak{B}(\mathcal{U})$ satisfies independence if the intial measure on $\mathcal{A}(\mathcal{U})$ satisfies independence. Let $\tilde{P}_{U|Y,X,Z}$ be the extension of $P_{U|Y,X,Z}$ from the proof of Theorem 3.1. Then for any $A\in\mathfrak{B}(\mathcal{U})$:

$$\begin{aligned}
&\tilde{P}_{U|Z}(A\mid Z=z_k)\\
&=\sum_{y\in\{0,1\}}\sum_{x\in\mathcal{X}}\sum_{s\in\{0,1\}^m}\mathbb{1}(A,\theta,s)P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta)\mid Y=y,X=x,Z=z_k\right)P_{Y,X|Z}(Y=y,X=x\mid Z=z_k)\\
&=\sum_{s\in\{0,1\}^m}\mathbb{1}(A,\theta,s)\sum_{y\in\{0,1\}}\sum_{x\in\mathcal{X}}P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta)\mid Y=y,X=x,Z=z_k\right)P_{Y,X|Z}(Y=y,X=x\mid Z=z_k)\\
&=\sum_{s\in\{0,1\}^m}\mathbb{1}(A,\theta,s)\sum_{y\in\{0,1\}}\sum_{x\in\mathcal{X}}P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta)\mid Y=y,X=x,Z=z_{k+1}\right)P_{Y,X|Z}(Y=y,X=x\mid Z=z_{k+1})\\
&=\sum_{y\in\{0,1\}}\sum_{x\in\mathcal{X}}\sum_{s\in\{0,1\}^m}\mathbb{1}(A,\theta,s)P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta)\mid Y=y,X=x,Z=z_{k+1}\right)P_{Y,X|Z}(Y=y,X=x\mid Z=z_{k+1})\\
&=\tilde{P}_{U|Z}(A\mid Z=z_{k+1}),
\end{aligned}$$

for all pairs z_k and z_{k+1} assigned positive probability, where the third equality follows from (B.2). Conclude that $\tilde{P}_{U|Z}$ satisfies the second condition in Definition B.1. ■

Analogous to Theorem 2.1, the first part of Corollary B.1 provides the theoretical link between the identified set for counterfactual conditional distributions and the identified set for the pair $(P_{U|Y,X,Z},\theta)$ under the additional independence assumption between U and Z. Furthermore, analogous to the result in Theorem 3.1, the second part of Corollary B.1 reduces an infinite dimensional existence problem to a finite dimensional existence problem. Importantly, the second part of Corollary B.1 builds on Theorem 3.1 by demonstrating that Assumption 4.1—which requires $P_{U|Z}(A \mid Z = z) = P_U(A)$ a.s. for all Borel sets A—can be imposed by considering only a finite number of equality constraints on a distribution $P_{U|Y,X,Z}$ defined on sets of the form $\mathcal{U}(s, \theta)$.

We have the following Corollary to Theorem 3.2:

Corollary B.2. *Under Assumptions 2.1, 2.2, and 4.1, the identified set for the counterfactual conditional probability $P_{Y_\gamma|Y,X,Z}(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j)$ is given by:*

$$\bigcup_{\theta\in\Theta} [\pi_{\ell b}(y, x_j, z_j, \theta), \pi_{ub}(y, x_j, z_j, \theta)]$$

where $\pi_{\ell b}(y, x_j, z_j, \theta)$ and $\pi_{ub}(y, x_j, z_j, \theta)$ are determined by the optimization problems:

$$\pi_{\ell b}(y, x_j, z_j, \theta) := \min_{\pi(\theta)\in\mathbb{R}^{d_\pi}} \sum_{s\in S_{\gamma(j)}} \pi(y, x_j, z_j, s, \theta), \text{ s.t. } (3.9), (3.10), (3.11), \text{ and } (4.1), \tag{B.3}$$

$$\pi_{ub}(y, x_j, z_j, \theta) := \max_{\pi(\theta)\in\mathbb{R}^{d_\pi}} \sum_{s\in S_{\gamma(j)}} \pi(y, x_j, z_j, s, \theta), \text{ s.t. } (3.9), (3.10), (3.11), \text{ and } (4.1). \tag{B.4}$$

Note that this Corollary is identical to Theorem 3.2 with the exception that we have imposed Assumption 4.1, and thus have included constraints of the form (4.1). With the exception of these additional constraints, the optimization problems that characterize the bounding problem are the same as before. Again, this result can be easily modified to bound any linear function of counterfactual conditional distributions by simply modifying the objective function in the optimization problems (B.3) and (B.4).

B.2 Monotonicity Assumptions

When we entertain Assumption 4.2, we have the following definition of the identified set, which is analogous to both Definitions 2.1 and 2.2.

Definition B.2. *Under Assumptions 2.1 and 4.2, the identified set $\mathcal{I}^*_{Y,X,Z}$ is the set of all pairs $(P_{U|Y,X,Z}, \theta)$ such that:*

(i) $(P_{U|Y,X,Z}, \theta)$ satisfies:

$$P_{U|Y,X,Z}(U \in \mathcal{U}(Y, X, Z, \theta) \mid Y = y, X = x, Z = z) = 1,$$

$P_{Y,X,Z}-$a.s.; and

(ii) For all $(j, k) \in \mathcal{M}$ from Assumption 4.2, we have:

$$P_{U|Y,X,Z}(\varphi(x_j, z_j, \theta, U) \leq \varphi(x_k, z_k, \theta, U) \mid Y = y, X = x, Z = z) = 1 \text{ a.s.}$$

Furthermore, under Assumptions 2.1, 2.2, and 4.2, the identified set of counterfactual conditional distribu-

tions $\mathcal{P}^*_{Y_\gamma|Y,X,Z,U}$ is the set of all conditional distributions $P_{Y_\gamma|Y,X,Z,U}$ satisfying:

$$P_{Y_\gamma|X,Z,U}\left(Y_\gamma = \mathbb{1}\{\varphi(\gamma(X,Z),U,\theta) \geq 0\} \mid Y = y, X = x, Z = z, U = u\right) = 1,$$

$P_{Y,X,Z,U}-a.s.$ for some pair $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$.

Again this definition and the results to follow are easily modified to accommodate the case when any combination of Assumptions 2.1 and 4.1 hold. We now provide the following Corollary whose proof follows almost identically to that of Theorems 2.1 and 3.1, with the exception being that we require condition (ii) of Definition B.2 to hold.

Corollary B.3. *Under Assumptions 2.1, 2.2, and 4.2, a counterfactual conditional distribution $P_{Y_\gamma|Y,X,Z}$ satisfies $P_{Y_\gamma|Y,X,Z} \in \mathcal{P}^*_{Y_\gamma|Y,X,Z}$ if and only if there exists a pair $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$ (for $\mathcal{I}^*_{Y,X,Z}$ from Definition B.2) satisfying:*

$$P_{Y_\gamma|X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x, Z = z\right) = P_{U|Y,X,Z}\left(\varphi(\gamma(X,Z),U,\theta) \geq 0 \mid Y = y, X = x, Z = z\right), \tag{B.5}$$

$P_{Y,X,Z}-a.s.$ Furthermore, for any collection of counterfactual conditional distributions $P_{Y_\gamma|Y,X,Z}$, there exists a collection of Borel conditional probability measures $P_{U|Y,X,Z}$ satisfying (B.5) with $(P_{U|Y,X,Z},\theta) \in \mathcal{I}^*_{Y,X,Z}$ (for $\mathcal{I}^*_{Y,X,Z}$ from Definition B.2) if and only if there exists a collection $P_{U|Y,X,Z}$ of probability measures on the sets in $\mathcal{A}(\theta)$ from (3.4) satisfying:

$$\sum_{s\in S_j} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = 1, X = x_j, Z = z_j\right) = 1,$$

$$\sum_{s\in S_j^c} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = 0, X = x_j, Z = z_j\right) = 1,$$

$$\sum_{s\in S_{\gamma(j)}} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x_j, Z = z_j\right) = P_{Y_\gamma|Y,X,Z}\left(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j\right),$$

for $y \in \{0,1\}$ and $j \in \{1,\dots,m\}$ assigned positive probability, and:

$$\sum_{s\in S_M^c} P_{U|Y,X,Z}\left(\mathcal{U}(s,\theta) \mid Y = y, X = x, Z = z\right) = 0, \;\; a.s. \tag{B.6}$$

for all (y,x,z) assigned positive probability, where S_M is as defined in Section 4.

The proof of this corollary is identical to the proof of Theorem 2.1 and Theorem 3.1. Analogous to Theorem 2.1, the first part of Corollary B.3 provides the theoretical link between the identified set for counterfactual conditional distributions and the identified set for the pair $(P_{U|Y,X,Z},\theta)$ under the additional monotonicity assumption. Analogous to Theorem 3.1, the second part of Corollary B.3 reduces an infinite dimensional existence problem to a finite dimensional existence problem amenable to analysis using optimization problems. Building on the intuition provided in example 2, the second part of Corollary B.3 demonstrates that monotonicity as in Assumption 4.2 can be imposed by considering only a finite number of equality constraints on a distribution $P_{U|Y,X,Z}$ defined on sets of the form $\mathcal{U}(s,\theta)$. By definition of the set

S_M, condition (B.6) simply assigns probability zero to all sets $\mathcal{U}(s,\theta)$ that do not satisfy the monotonicity relation from Assumption 4.2. This leads to the following result.

Corollary B.4. *Under Assumptions 2.1, 2.2, and 4.2, the identified set for the counterfactual conditional probability $P_{Y_\gamma|Y,X,Z}(Y_\gamma = 1 \mid Y = y, X = x_j, Z = z_j)$ is given by:*

$$\bigcup_{\theta\in\Theta} [\pi_{\ell b}(y,x_j,z_j,\theta), \pi_{ub}(y,x_j,z_j,\theta)]$$

where $\pi_{\ell b}(y,x_j,z_j,\theta)$ and $\pi_{ub}(y,x_j,z_j,\theta)$ are determined by the optimization problems:

$$\pi_{\ell b}(y,x_j,z_j,\theta) := \min_{\pi(\theta)\in\mathbb{R}^{d_\pi}} \sum_{s\in S_{\gamma(j)}} \pi(y,x_j,z_j,s,\theta), \text{ s.t. (3.9), (3.10), (3.11), and (4.2),} \tag{B.7}$$

$$\pi_{ub}(y,x_j,z_j,\theta) := \max_{\pi(\theta)\in\mathbb{R}^{d_\pi}} \sum_{s\in S_{\gamma(j)}} \pi(y,x_j,z_j,s,\theta), \text{ s.t. (3.9), (3.10), (3.11), and (4.2).} \tag{B.8}$$

Note that this Corollary is identical to Theorem 3.2 with the exception that we have imposed Assumption 4.2, and thus have included constraints of the form (4.2). With the exception of these additional constraints, the optimization problems that characterize the bounding problem are the same as before. Finally, alternative counterfactual quantities can be bounded in the same way by simply modifying the objective function in (B.7) and (B.8).

B.3 Consistency

In this subsection we present a basic consistency result for functionals of a partially identified parameter. The result is designed to minimize the number of high-level assumptions required for consistency, and is closely related to results found in Molchanov (1998), Manski and Tamer (2002), and Chernozhukov et al. (2007), possibly among others. It is presented in a form that is more general than necessary for the current paper, and so it may be of interest in other applications.

We consider an environment where the researcher wishes to compute bounds on a functional $\mathbb{E}_P[\psi(W_i,\tau_1,\tau_2)]$, where $\psi : \mathcal{W}\times\mathcal{T} \to \mathbb{R}$, where $\mathcal{W} \subset \mathbb{R}^{d_w}$ denotes the support of the observed random vector W, and $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \subset \mathbb{R}^{d_\tau}$ denotes the parameter space with typical elements $\tau = (\tau_1,\tau_2) \in \mathcal{T}$. The values of (τ_1,τ_2) are constrained by J moment inequalities of the form:

$$\mathbb{E}_P[m_j(W_i,\tau_1,\tau_2)] \leq 0, \text{ for } j = 1,\ldots,J.$$

Note this does not rule out moment equalities, since each moment equality can be equivalently written as a combination of two moment inequalities. In this environment, the identified set for $(\tau_{01},\tau_{02}) \in \mathcal{T}$ at the true P is given by:

$$\mathcal{T}^*(P) := \{(\tau_1,\tau_2) \in \mathcal{T} : \mathbb{E}_P[m_j(W_i,\tau_1,\tau_2)] \leq 0 \text{ for } j = 1,\ldots,J\}.$$

In addition, the identified set for $\psi_0 := \mathbb{E}_P[\psi(W_i, \tau_{01}, \tau_{02})]$ is given by:

$$\Psi^*(P) := \left\{\overline{\psi} \in \mathbb{R} : \exists(\tau_1, \tau_2) \in \mathcal{T}_I(P) \text{ s.t. } \overline{\psi} = \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)]\right\}.$$

Let us define the projection:

$$\mathcal{T}_1^*(\tau_2, P) := \{\tau_1 \in \mathcal{T}_1 : \mathbb{E}_P[m_j(W_i, \tau_1, \tau_2)] \leq 0 \text{ for } j = 1, \ldots, J\}.$$

It is then straightforward to show that $\Psi^*(P)$ can be rewritten as:

$$\Psi^*(P) = \bigcup_{\tau_2 \in \mathcal{T}_2} [\Psi_{\ell b}(\tau_2, P), \Psi_{ub}(\tau_2, P)],$$

where:

$$\Psi_{\ell b}(\tau_2, P) := \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)], \qquad \Psi_{ub}(\tau_2, P) := \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)].$$

We study the consistency properties of the sample analog estimator for this representation of $\Psi^*(P)$. In particular, define:

$$\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] := \frac{1}{n}\sum_{i=1}^n \psi(W_i, \tau_1, \tau_2), \quad \mathbb{E}_n[m_j(W_i, \tau_1, \tau_2)] := \frac{1}{n}\sum_{i=1}^n m_j(W_i, \tau_1, \tau_2), \text{ for } j = 1, \ldots, J.$$

Then the sample analog estimator of interest is given by:

$$\Psi^*(\mathbb{P}_n) = \bigcup_{\tau_2 \in \mathcal{T}_2} [\Psi_{\ell b}(\tau_2, \mathbb{P}_n), \Psi_{ub}(\tau_2, \mathbb{P}_n)],$$

where:

$$\Psi_{\ell b}(\tau_2, \mathbb{P}_n) := \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, \mathbb{P}_n)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)], \qquad \Psi_{ub}(\tau_2, \mathbb{P}_n) := \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, \mathbb{P}_n)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)],$$

and:

$$\mathcal{T}_1^*(\tau_2, \mathbb{P}_n) := \{\tau_1 \in \mathcal{T}_1 : \mathbb{E}_n[m_j(W_i, \tau_1, \tau_2)] \leq 0 \text{ for } j = 1, \ldots, J\}.$$

In the following, we define the sequence $\{\eta_n(\tau_2)\}_{n=1}^\infty$ as:

$$\eta_n(\tau_2) := \max\left\{\max_{j=1,\ldots,J.} \sup_{\tau_1 \in \mathcal{T}_1} |\mathbb{E}_n[m_j(W_i, \tau_1, \tau_2)] - \mathbb{E}_P[m_j(W_i, \tau_1, \tau_2)]|, \sup_{\tau_1 \in \mathcal{T}_1} |\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)]|\right\}.$$

We impose the following assumption.

Assumption B.1. *(i) The parameter space $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \subset \mathbb{R}^{d_\tau}$, where $\mathcal{T}_1$ is compact; (ii) for each $\tau_2 \in \mathcal{T}_2$, the function $\psi(\,\cdot\,, \tau_2) : \mathcal{W} \times \mathcal{T}_1 \to \mathbb{R}$ is measurable in $W_i \in \mathcal{W} \subset \mathbb{R}^{d_w}$ and is Lipschitz continuous in τ_1 with a (possibly data-dependent) Lipschitz constant $C(\tau_2)$ with $\sup_{\tau_2 \in \mathcal{T}_2} C(\tau_2) < \infty$ a.s.; (iii) for $j = 1, \ldots, J$, and for each $\tau_2 \in \mathcal{T}_2$, the moment function $m_j(\,\cdot\,, \tau_2) : \mathcal{W} \times \mathcal{T}_1 \to \mathbb{R}$ is measurable in W_i and lower semicontinuous in τ_1; (iv) the true data generating process is indexed by a triple $(\tau_{01}, \tau_{02}, P)$ that satisfies $(\tau_{01}, \tau_{02}) \in \mathcal{T}$,* *and* $\mathbb{E}_P[m_j(W_i,\tau_{01},\tau_{02})] \leq 0$*, for* $j = 1,\ldots,J$*; (v) the sample* $\{W_i\}_{i=1}^n$ *is an indepndent and identically distributed draw from* P*; (vi) for each fixed* $\tau_2 \in \mathcal{T}_2$*, we have* $\eta_n(\tau_2) = O_P(a_n^{-1})$ *for some sequence* $a_n \uparrow \infty$*; (vii) for each fixed* $\tau_2 \in \mathcal{T}$*, there exists a sequence* $b_n \downarrow 0$ *satisfying* $b_n \geq \eta_n(\tau_2)$ *with probability approaching 1 (w.p.a. 1).; (viii) there exists a finite subset* $\mathcal{T}_2' \subset \mathcal{T}_2$ *such that:*

$$\{\tau_1 \in \mathcal{T}_1 : \exists \tau_2 \in \mathcal{T}_2 \textit{ s.t. } \mathbb{E}_P[m_j(W_i,\tau_1,\tau_2)] \leq 0 \textit{ for } j = 1,\ldots,k\}$$
$$= \{\tau_1 \in \mathcal{T}_1 : \exists \tau_2 \in \mathcal{T}_2' \textit{ s.t. } \mathbb{E}_P[m_j(W_i,\tau_1,\tau_2)] \leq 0 \textit{ for } j = 1,\ldots,k\}\,.$$

Part (i) of Assumption B.1 is standard in the literature on extremum estimators. Part (ii) separates the roles of τ_1 and τ_2, and restricts the objective function to be Lipschitz continuous in the parameter τ_1 for each τ_2. Part (ii) places no restrictions on how τ_2 enters the objective function. Part (iii) further separates the roles of τ_1 and τ_2 by requiring each of the moment functions to be lower semicontinuous in τ_1. Similar to part (ii), no restrictions are placed on how τ_2 enters the moment functions. Assumption (iv) is standard, and simply indicates that the true parameters satisfying the moment inequalities at the true P. Part (v) is also standard, although it rules out the case of dependent data. Part (vi) indicates that $\eta_n(\tau_2)$ converges in probability at a rate of $1/a_n$. This can be verified using standard assumptions; for example, if for each $\tau_2 \in \mathcal{T}_2$ the $J+1$ classes of functions:

$$\mathcal{F}_\psi(\tau_2) := \{\psi(\,\cdot\,,\tau_1,\tau_2) : \mathcal{W} \to \mathbb{R} \mid \tau_1 \in \mathcal{T}_1\},$$
$$\mathcal{F}_j(\tau_2) := \{m_j(\,\cdot\,,\tau_1,\tau_2) : \mathcal{W} \to \mathbb{R} \mid \tau_1 \in \mathcal{T}_1\}, \text{ for } j = 1,\ldots,J,$$

are all $P-$Donsker classes, then part (vi) is satisfied with $a_n = \sqrt{n}$. This is the case, for example, for all specifications considered in Section 5. After verifying part (vi), it is easy to find a sequence b_n satisfying part (vii). For example, if $a_n = \sqrt{n}$ from part (vi), then we can set $b_n = b/\sqrt{\log(n)}$ for any $b > 0$. Finally, part (viii) essentially allows us to replace $\mathcal{T}_2$ with a finite subset $\mathcal{T}_2'$ without impacting the bounding problem. It is precisely because of part (viii) that all other parts of Assumption B.1—namely parts (ii), (iii), (vi) and (vii)—are allowed to be so flexible with respect to the parameter τ_2. This last component of Assumption B.1 is verified in our basic setup in Proposition 3.1 in Gu and Russell (2021), the previous version of this paper. Gu and Russell (2021) also verify the assumption under the functional form, independence, and monotonicity assumptions discussed in the main text. All other components of Assumption B.1 are either standard assumptions, or are easily verified for the bounding problems presented in the main text and for all specifications considered in Section 5.

Before stating the main result for this subsection, for any $c \in \mathbb{R}$ let us define:

$$\mathcal{T}_1^*(\tau_2,P,c) := \{\tau_1 \in \mathcal{T}_1 : \mathbb{E}_P[m_j(W,\tau_1,\tau_2)] \leq c \text{ for } j = 1,\ldots,J\}\,,$$

and:

$$\Psi^*(P,c) = \bigcup_{\tau_2 \in \mathcal{T}_2'} [\Psi_{\ell b}(\tau_2, P, c), \Psi_{ub}(\tau_2, P, c)],$$

where:

$$\Psi_{\ell b}(\tau_2, P, c) := \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P, c)} \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)], \qquad \Psi_{ub}(\tau_2, \mathbb{P}_n, c) := \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P, c)} \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)].$$

Define the sets $\mathcal{T}_1^*(\tau_2, P, c)$ and $\Psi^*(P,c)$ analogously. The following Theorem then shows that a slight enlargement of the set $\Psi^*(\mathbb{P}_n)$ is a consistent estimator for the set $\Psi^*(P)$, where consistency is defined using the Hausdorff metric.

Theorem B.1. *Suppose that Assumption B.1 holds. Then $d_H(\Psi^*(\mathbb{P}_n, b_n), \Psi^*(P)) = o_P(1)$, where b_n is the sequence from Assumption B.1.*

Proof of Theorem B.1. We have:

$$d_H(\Psi^*(\mathbb{P}_n, b_n), \Psi^*(P)) \leq \sum_{\tau_2 \in \mathcal{T}_2'} d_H\left([\Psi_{\ell b}(\tau_2, \mathbb{P}_n, b_n), \Psi_{ub}(\tau_2, \mathbb{P}_n, b_n)], [\Psi_{\ell b}(\tau_2, P), \Psi_{ub}(\tau_2, P)]\right).$$

Since $\mathcal{T}_2'$ is finite by Assumption B.1(viii), it suffices to show that:

$$d_H\left([\Psi_{\ell b}(\tau_2, \mathbb{P}_n, b_n), \Psi_{ub}(\tau_2, \mathbb{P}_n, b_n)], [\Psi_{\ell b}(\tau_2, P), \Psi_{ub}(\tau_2, P)]\right) = o_P(1),$$

for each $\tau_2 \in \mathcal{T}_2'$. To this end, fix any $\tau_2 \in \mathcal{T}_2$. To show the previous display, it suffices to show consistency of the upper and lower bounds; i.e. that $|\Psi_{\ell b}(\tau_2, \mathbb{P}_n, b_n) - \Psi_{\ell b}(\tau_2, P)| = o_P(1)$ and that $|\Psi_{ub}(\tau_2, \mathbb{P}_n, b_n) - \Psi_{ub}(\tau_2, P)| = o_P(1)$. We focus on the lower bound, since the upper bound proof is symmetric.

First recall that $\psi(W_i, \tau_1, \tau_2)$ is continuous with respect to τ_1 for every τ_2 by Assumption B.1(ii), and $\mathcal{T}_1$ is compact by Assumption B.1(i). Thus, we have that $\psi(W_i, \tau_1, \tau_2)$ is uniformly continuous (w.r.t. τ_1) on $\mathcal{T}_1$. Thus, for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that $|\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \mathbb{E}_n[\psi(W_i, \tau_1', \tau_2)]| < \varepsilon$ whenever $||\tau_1 - \tau_1'|| < \delta(\varepsilon)$. Now note that:

$$\begin{aligned}
&|\Psi_{\ell b}(\tau_2, \mathbb{P}_n, b_n) - \Psi_{\ell b}(\tau_2, P)| \\
&= \left| \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_P[\psi(W, \tau_1, \tau_2)] \right|, \\
&\leq \left| \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] \right| \\
&\qquad + \left| \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \min_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} \mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)] \right|, \\
&= \left| \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} -\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n)} -\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] \right| \\
&\qquad + \left| \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} -\mathbb{E}_P[\psi(W_i, \tau_1, \tau_2)] - \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2, P)} -\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] \right|, \\
&\leq \max_{\{\tau_1, \tau_1' \in \mathcal{T}_1 : ||\tau_1 - \tau_1'|| \leq d_H(\mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n), \mathcal{T}_1^*(\tau_2, P))\}} |-\mathbb{E}_n[\psi(W_i, \tau_1, \tau_2)] - -\mathbb{E}_n[\psi(W_i, \tau_1', \tau_2)]|
\end{aligned}$$

$$
\begin{aligned}
&\quad + \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2,P)} \left|-\mathbb{E}_n[\psi(W_i,\tau_1,\tau_2)] - -\mathbb{E}_P[\psi(W_i,\tau_1,\tau_2)]\right| \\
&\leq \max_{\{\tau_1,\tau_1' \in \mathcal{T}_1 : ||\tau_1-\tau_1'|| \leq d_H(\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n),\mathcal{T}_1^*(\tau_2,P))\}} \left|\mathbb{E}_n[\psi(W_i,\tau_1',\tau_2)] - \mathbb{E}_n[\psi(W_i,\tau_1,\tau_2)]\right| \\
&\quad + \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2,P)} \left|\mathbb{E}_P[\psi(W_i,\tau_1,\tau_2)] - \mathbb{E}_n[\psi(W_i,\tau_1,\tau_2)]\right| \\
&= \max_{\{\tau_1,\tau_1' \in \mathcal{T}_1 : ||\tau_1-\tau_1'|| \leq d_H(\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n),\mathcal{T}_1^*(\tau_2,P))\}} C \cdot ||\tau_1 - \tau_1'|| + \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2,P)} \left|\mathbb{E}_P[\psi(W_i,\tau_1,\tau_2)] - \mathbb{E}_n[\psi(W_i,\tau_1,\tau_2)]\right| \\
&\leq C \cdot d_H(\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n),\mathcal{T}_1^*(\tau_2,P)) + \max_{\tau_1 \in \mathcal{T}_1^*(\tau_2,P)} \left|\mathbb{E}_P[\psi(W_i,\tau_1,\tau_2)] - \mathbb{E}_n[\psi(W_i,\tau_1,\tau_2)]\right|.
\end{aligned}
$$

It suffices to show the two terms in the last line of the previous display converge to zero in probability. The second term converges in probability to zero by Assumption B.1(vi). Furthermore, since $C < \infty$ w.p. 1, the first term converges to zero in probability if we can show that:

$$
d_H(\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n),\mathcal{T}_1^*(\tau_2,P)) = o_P(1).
$$

The remainder of the proof focuses on proving this latter fact. Note that:

$$
d_H(\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n),\mathcal{T}_1^*(\tau_2,P)) = \inf\{\delta > 0 : \mathcal{T}_1^*(\tau_2,P) \subseteq \mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n)^\delta, \text{ and } \mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n) \subseteq \mathcal{T}_1^*(\tau_2,P)^\delta\},
$$

where:

$$
\begin{aligned}
\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n)^\delta &:= \{\tau_1 \in \mathcal{T}_1 : B_\delta(\tau_1) \cap \mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n) \neq \varnothing\}, \\
\mathcal{T}_1^*(\tau_2,P)^\delta &:= \{\tau_1 \in \mathcal{T}_1 : B_\delta(\tau_1) \cap \mathcal{T}_1^*(\tau_2,P) \neq \varnothing\},
\end{aligned}
$$

where $B_\delta(\tau_1)$ denotes the closed ball of radius $\delta > 0$ around τ_1. The next part of the proof closely follows the proof of Theorem 2.1 in Molchanov (1998). Define the function:

$$
\rho(\varepsilon) := d_H(\mathcal{T}_1^*(\tau_2,P,\varepsilon),\mathcal{T}_1^*(\tau_2,P)).
$$

Since each of the moment functions are lower semi-continuous in τ_1 for each τ_2, each of the sets $\mathcal{T}_1^*(\tau_2,P,\varepsilon)$ and $\mathcal{T}_1^*(\tau_2,P)$ are closed and ρ is right continuous. Furthermore, ρ is non-increasing for $\varepsilon < 0$ and non-decreasing for $\varepsilon > 0$. Now by Assumption B.1 we have with high probability:

$$
\begin{aligned}
\mathcal{T}_1^*(\tau_2,\mathbb{P}_n,b_n) &= \{\tau_1 \in \mathcal{T}_1 : \mathbb{E}_n[m_j(W,\tau_1,\tau_2)] \leq b_n \text{ for } j = 1,\ldots,k\} \\
&\subseteq \{\tau_1 \in \mathcal{T}_1 : \mathbb{E}_n[m_j(W,\tau_1,\tau_2)] \leq \eta_n(\tau_2) + b_n \text{ for } j = 1,\ldots,k\} \\
&\subseteq \mathcal{T}_1^*(\tau_2,P,2b_n) \\
&\subseteq \mathcal{T}_1^*(\tau_2,P)^{\rho(2b_n)}.
\end{aligned}
$$

Furthermore, by Assumption B.1 we have with high probability for large enough n:

$$
\mathcal{T}_1^*(\tau_2,P) \subseteq \mathcal{T}_1^*(\tau_2,P,b_n - \eta_n(\tau_2))
$$

$$\subseteq \mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n).$$

Conclude that with high probability for large enough n:

$$d_H(\mathcal{T}_1^*(\tau_2, \mathbb{P}_n, b_n), \mathcal{T}_1^*(\tau_2, P)) \leq \rho(2b_n) \to 0,$$

where the last line follows from right-continuity of the function $\rho(\cdot)$. Since $\tau_2 \in \mathcal{T}_2'$ was arbitrary, this completes the proof. ■

B.4 Bias-Corrected Estimates and Inference

In Section 5 we use the inference method of Cho and Russell (2020), which is specifically designed for uniform inference on value functions in stochastic linear programming problems. However, the characterization of the identified set provided by Theorem 3.2 is slightly different then the setting considered in Cho and Russell (2020). In particular, the identified set in Theorem 3.2 is a union of intervals whose endpoints are determined by the value functions of two linear programming problems.

To extend the result of Cho and Russell (2020), let ψ_0 denote the true value of our counterfactual object of interest (e.g. a counterfactual conditional choice probability) and let $\Psi^*(P)$ denote the identified set for ψ_0 evaluated at a distribution P belonging to some class of distributions $\mathcal{P}$ characterized by Assumption 3.2 in Cho and Russell (2020). For some $\alpha \in (0,1)$, we would like to construct a random set $CS_n(1-\alpha)$ satisfying:

$$\liminf_{n\to\infty} \inf_{\{(\psi,P):\psi\in\Psi^*(P), P\in\mathcal{P}\}} \text{Pr}_P\left(\psi_0 \in CS_n(1-\alpha)\right) \geq 1-\alpha.$$

Let $\Psi^*(\theta, P) := [\psi_{\ell b}(\theta, P), \psi_{ub}(\theta)]$ where $\psi_{\ell b}(\theta, P)$ is the value function from (3.13) and $\psi_{ub}(\theta, P)$ is the value function from (3.14) for some distribution $P \in \mathcal{P}$. Then from Theorem 3.2 we have that:

$$\Psi^*(P) = \bigcup_{\theta\in\Theta} \Psi^*(\theta, P). \tag{B.9}$$

By Proposition 3.2, there exists a finite set $\Theta' \subseteq \Theta$ of representative points satisfying:

$$\Psi^*(P) = \bigcup_{\theta\in\Theta'} \Psi^*(\theta, P).$$

Now consider setting:

$$CS_n(1-\alpha) = \bigcup_{\theta\in\Theta'} CS_n(1-\alpha, \theta),$$

where the random sets $\{CS_n(1-\alpha,\theta) : \theta \in \Theta'\}$ satisfy:

$$\liminf_{n\to\infty} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P), P\in\mathcal{P}\}} \text{Pr}_P\left(\psi_0 \in CS_n(1-\alpha, \theta)\right) \geq 1-\alpha. \tag{B.10}$$

Then combining everything we can write:

$$\begin{aligned}
&\liminf_{n\to\infty} \inf_{\{(\psi,P):\psi\in\Psi^*(P),P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in CS_n(1-\alpha)\right) \\
&= \liminf_{n\to\infty} \inf_{\{(\psi,P):\psi\in\Psi^*(P,\theta),\theta\in\Theta',P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in CS_n(1-\alpha)\right) \\
&= \liminf_{n\to\infty} \min_{\theta\in\Theta'} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P),P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in CS_n(1-\alpha)\right) \\
&= \liminf_{n\to\infty} \min_{\theta\in\Theta'} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P),P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in \bigcup_{\theta\in\Theta'} CS_n(1-\alpha,\theta)\right) \\
&\geq \liminf_{n\to\infty} \min_{\theta\in\Theta'} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P),P\in\mathcal{P}\}} \min_{\theta\in\Theta'} \Pr_P\left(\psi_0 \in CS_n(1-\alpha,\theta)\right) \\
&= \liminf_{n\to\infty} \min_{\theta\in\Theta'} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P),P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in CS_n(1-\alpha,\theta)\right) \\
&= \min_{\theta\in\Theta'} \liminf_{n\to\infty} \inf_{\{(\psi,P):\psi\in\Psi^*(\theta,P),P\in\mathcal{P}\}} \Pr_P\left(\psi_0 \in CS_n(1-\alpha,\theta)\right) \\
&\geq 1-\alpha,
\end{aligned}$$

where the second last line follows from continuity of the minimum, and the last line follows from (B.10). Thus, it suffices to construct random sets $CS_n(1-\alpha,\theta)$ satisfying (B.10) for each $\theta \in \Theta'$. Since in all specifications in the application section the representative points are known, the confidence sets $CS_n(1-\alpha,\theta)$ are constructed for each representative point using the procedure in Cho and Russell (2020), and our final confidence set is given by (B.9). After introducing additional moment assumptions on the random variables in our application, Assumptions 3.1 and 3.2 in Cho and Russell (2020) (the only two assumptions required for their method) are easily verified.

Finally, in the application in Section 5, we report bias-corrected estimates of the upper and lower endpoints of the identified set. In particular, if $\psi_{\ell b}(P)$ is the lower endpoint of the (convex hull of the) identified set $\Psi^*(P)$ and $\psi_{ub}(P)$ is the upper endpoint of the (convex hull of the) identified set $\Psi^*(P)$, then our estimates $\hat{\psi}_{\ell b}$ and $\hat{\psi}_{ub}$ are half-median unbiased in the sense that $\hat{\psi}_{\ell b} \leq \psi_{\ell b}(P)$ and $\psi_{ub}(P) \leq \hat{\psi}_{ub}$, both holding with probability at least 1/2 uniformly over $P \in \mathcal{P}$. The use of half-median unbiased estimators was proposed by Chernozhukov et al. (2013). In our case, these bias-corrected estimates can also be constructed using the inference procedure of Cho and Russell (2020). In particular, Cho and Russell (2020) show how to construct one-sided confidence intervals, and the procedure discussed above is easily amended for the one-sided case. The estimates of the identified set reported in the application in Section 5 are the resulting $\alpha = 0.5$ one-sided lower and upper confidence bounds.

B.5 The Additively Separable Case

In this subsection we show how our method can be applied to a model that satisfies the following assumption.

Assumption B.2. *The index function φ satisfying Assumption 2.1 is additively separable in U; i.e. we have $\varphi(X,Z,U,\theta) = \tilde{\varphi}(X,Z,\theta) - U$ for some function $\tilde{\varphi}$.*

This is a well-studied special case of the linear model considered in the main text. In particular, much of the discussion in this section expands upon the insights of Chesher (2013). We consider two cases: (i) when the structural function φ is linear in the parameter vector θ, and (ii) when the structural function is unknown. To begin, let us consider the following simple example.

Example 3. *Suppose we have a scalar variable X with support $\mathcal{X} = \{x_1, \ldots, x_{m_x}\}$ and latent variables $U \in [-1, 1]$, and suppose there are no variables Z. Consider the following additively separable threshold crossing model:*

$$Y = \mathbb{1}\{X\theta \geq U\},$$

where θ is a fixed scalar coefficient. The response types in this setting are characterized by the $m_x \times 1$ vectors:

$$r(u, \theta) := \begin{bmatrix} \mathbb{1}\{x_1\theta \geq u\} \\ \mathbb{1}\{x_2\theta \geq u\} \\ \vdots \\ \mathbb{1}\{x_{m_x}\theta \geq u\} \end{bmatrix}.$$

However, the set of possible response types in this setting depends on the sign of the fixed coefficient θ. In particular, when $\theta \geq 0$ we have the response types $r(u, \theta) \in \{s_1, \ldots, s_{m_x+1}\}$, where:

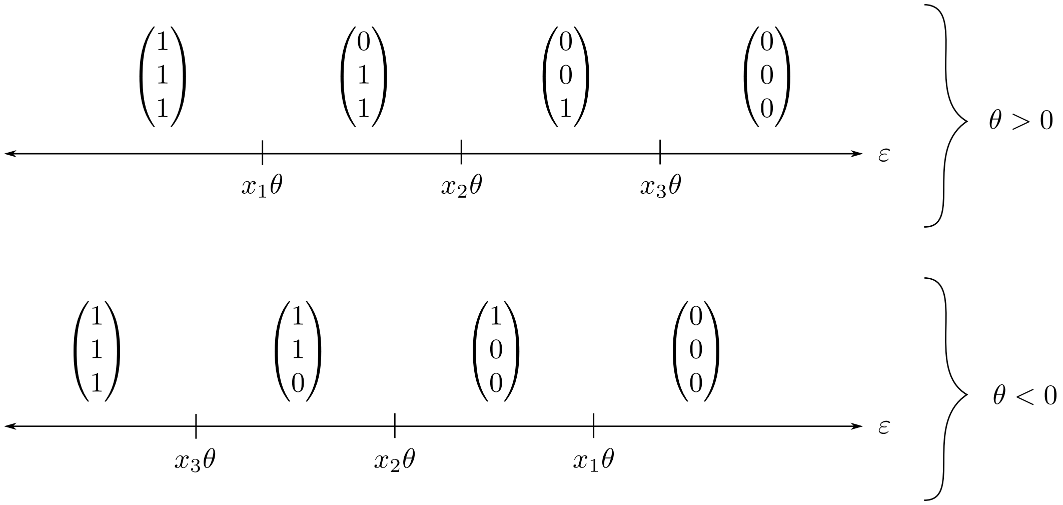
$$s_1 := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \qquad s_2 := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \qquad \ldots, \qquad s_{m_x} := \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, \qquad s_{m_x+1} := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}. \tag{B.11}$$

No other response types are possible when $\theta > 0$, and so all other response types must be assigned zero probability. Alternatively, when $\theta < 0$ we have the response types $r(u, \theta) \in \{s_1', \ldots, s_{m_x+1}'\}$, where:

$$s_1' := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \qquad s_2' := \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \qquad \ldots, \qquad s_{m_x}' := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \qquad s_{m_x+1}' := \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}. \tag{B.12}$$

Again, no other response types are possible when $\theta < 0$, and so must be assigned zero probability by the distribution of U.

The reason that these particular response types arise when $\theta \geq 0$ and $\theta < 0$ is due to the ordering of the support of X induced by the value of the scalar product $X\theta$. In particular, if we suppose $x_1 \leq x_2 \leq \ldots \leq x_{m_x}$, then when $\theta \geq 0$ we have the ordering $x_1\theta \leq x_2\theta \leq \ldots \leq x_{m_x}\theta$. This means, for example, that it is impossible



$\begin{pmatrix}1\\1\\1\end{pmatrix}$
$\begin{pmatrix}0\\1\\1\end{pmatrix}$
$\begin{pmatrix}0\\0\\1\end{pmatrix}$
$\begin{pmatrix}0\\0\\0\end{pmatrix}$
$x_1\theta$
$x_2\theta$
$x_3\theta$
ε
$\theta > 0$
$\begin{pmatrix}1\\1\\1\end{pmatrix}$
$\begin{pmatrix}1\\1\\0\end{pmatrix}$
$\begin{pmatrix}1\\0\\0\end{pmatrix}$
$\begin{pmatrix}0\\0\\0\end{pmatrix}$
$x_3\theta$
$x_2\theta$
$x_1\theta$
ε
$\theta < 0$

Figure 4: A figure corresponding to Example 3 illustrating the partition of the latent variable space according to response types in the case when the index function is additively separable in U and when $\mathcal{X} = \{x_1, x_2, x_3\}$ with $x_1 \leq x_2 \leq x_3$. As indicated in the example, the feasible response types are those that correspond to a particular ordering of the points in $\mathcal{X}$ induced by the scalar product $X\theta$.

to find a value of $u \in [-1, 1]$ so that:

$$r(u,\theta) = \begin{bmatrix} \mathbb{1}\{x_1\theta \geq u\} \\ \mathbb{1}\{x_2\theta \geq u\} \\ \mathbb{1}\{x_3\theta \geq u\} \\ \vdots \\ \mathbb{1}\{x_{m_x}\theta \geq u\} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Indeed, the existence of such a value for u would contradict the ordering $x_1\theta \leq x_2\theta \leq \ldots \leq x_{m_x}\theta$. This means that when $\theta \geq 0$ certain response types are not possible, and so must be assigned probability zero by the distribution of U. An identical intuition holds in the case when $\theta < 0$. In the end, the response types that can be assigned positive probability in this example when $\theta \geq 0$ and $\theta < 0$ are exactly the ones corresponding to the vectors in (B.11) *and* (B.12)*, respectively. Figure 4 provides an illustration in the case when $\mathcal{X} = \{x_1, x_2, x_3\}$.*

This example illustrates the key ideas behind the implementation of our approach when the index function is additively separable in U, as in Assumption B.2. In particular, given the function $\tilde{\varphi}$ from Assumption B.2, the key is to determine the values of θ such that the function $\tilde{\varphi}(\,\cdot\,, \theta) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ induces a unique ordering of the points in the support $\mathcal{X} \times \mathcal{Z}$. With no Z variables, a scalar X variable, and $\tilde{\varphi}(X, Z, \theta) = X\theta$, Example 3 shows that only two orderings are possible, corresponding to the case when $\theta \geq 0$ and $\theta < 0$. After the order is determined, we can immediately determine the set of response types that must be assigned zero probability by the distribution of U, and then impose these restrictions as an additional constraint in the bounding problems (3.13) and (3.14). In particular, letting S_φ denote the set of all binary vectors

$s \in \{0,1\}^m$ corresponding to sets $\mathcal{U}(s,\theta)$ that can be assigned positive probability under Assumption B.2, and impose the constraint:

$$\sum_{s \in S_{\varphi}^{c}} \pi(y, x_j, z_j, \theta, s) = 0, \tag{B.13}$$

for all $y \in \{0,1\}$ and $j = 1, \ldots, m$ occurring with positive probability. Thus, Theorem 3.2 can be extended to accommodate Assumption B.2 by simply adding the constraints (B.13) to the optimization problems (3.13) and (3.14).

Similar to the discussion in the main text, determining the sets $\mathcal{U}(s,\theta)$ that can be assigned positive probability under Assumption B.2 poses an interesting computational problem. Although Example 3 illustrates a case when there are only two orderings, in general many more orderings may be possible, even when $\tilde{\varphi}$ is linear in θ. Clearly at most $m!$ orderings are possible, but when the index function is linear in θ it is possible to show that the maximum number of possible orderings is much smaller than $m!$. In particular, partition $\theta = (\theta_x, \theta_z)$ and consider the function $\tilde{\varphi}(X, Z, \theta) = X\theta_x + Z\theta_z$ where X is a vector of dimension d_x and Z is a vector of dimension d_z. Label the support $\mathcal{X} \times \mathcal{Z}$ as $\{(x_1, z_1), (x_2, z_2), \ldots, (x_m, z_m)\}$, and let $\Delta_{jk} := (x_j, z_j) - (x_k, z_k)$ for $1 \leq j < k \leq m$. Setting $d = d_x + d_z$, the set $H_{jk} := \{\theta \in \mathbb{R}^d : \Delta_{jk}\theta = 0\}$ defines a hyperplane through the origin that is normal to the line connecting (x_j, z_j) and (x_k, z_k) in $\mathbb{R}^d$. The set of all such hyperplanes partitions $\mathbb{R}^d$ into at most $Q(m,d)$ non-empty cones, where $Q(m,d)$ is defined recursively as:

$$Q(m,d) = Q(m-1,d) + (m-1)Q(m-1,d-1), \tag{B.14}$$

with $Q(m,1) = 2$ for all $m \geq 2$ and $Q(2,d) = 2$ for all $d \geq 1$. Furthermore, each these non-empty cones corresponds exactly to the equivalence class of vectors $\theta = (\theta_x, \theta_z)$ that induce a unique ordering of the points in $\mathcal{X} \times \mathcal{Z}$. Thus, the value $Q(m,d)$ serves as an upper bound on the number of orderings of the points in $\mathcal{X} \times \mathcal{Z}$ that are inducible by the function $\tilde{\varphi}(X, Z, \theta) = X\theta_x + Z\theta_z$. The recursive formula from (B.14) defining the upper bound $Q(m,d)$ has been independently discovered in different contexts by many authors; the earliest such account appears in Bennett (1956), although the formula was independently discovered again in Cover (1967). The upper bound $Q(m,d)$ is obtained when the collection of hyperplanes of the form H_{jk} are in general position. Note that $Q(m,1) = 2$ corresponds exactly to Example 3, where it was shown that only two orderings could be induced when $\tilde{\varphi}(X, Z, \theta) = X\theta$ for scalar X and θ. Typically, $Q(m,d) < m!$, although some inspection of the formula shows that we always have $Q(m,d) = m!$ when $d \geq m - 1$.

If we could select one value of θ from each of the cones defined by the collection of hyperplanes of the form H_{jk}, we could then determine the permitted orderings of the support points $\mathcal{X} \times \mathcal{Z}$ by simply evaluating $x_j\theta_x + z_j\theta_z$ for $j = 1, \ldots, m$ at the selected value for θ. This would then allow us to determine which sets $\mathcal{U}(s,\theta)$ must be assigned zero probability under Assumption B.2. Note that under Assumption B.2 the latent variable U obtains a value on the hyperplane H_{jk} with probability zero. Thus, it suffices to select one value of θ from each of the *non-empty* cones defined by the collection of hyperplanes of the form H_{jk}. However, this can be done using the hyperplane arrangement algorithm described in the main text applied to the hyperplanes of the form H_{jk} for $1 \leq j < k \leq m$.

Our method is also applicable to cases when $\tilde{\varphi}(X, Z, \theta)$ may be non-linear in the finite-dimensional vector θ. To see how this case can be accommodated, recall that the case when $\tilde{\varphi}$ is linear in θ, the ordering of the support points in $\mathcal{X} \times \mathcal{Z}$ by the function $\tilde{\varphi}(X, Z, \theta)$ allowed us to determine the admissible response types, which in turn allowed us to construct the additional constraints needed in programs (3.13) and (3.14). A similar strategy can be used when $\tilde{\varphi}$ is not known by the researcher. However, when $\tilde{\varphi}$ is not restricted by the researcher, all orderings of the support points in $\mathcal{X} \times \mathcal{Z}$ are possible. The procedure to bound a counterfactual probability (or some other counterfactual quantity of interest) is then as follows. The researcher must first fix an ordering of the support points in $\mathcal{X} \times \mathcal{Z}$, determine the admissible response types S_{φ} for the fixed ordering, and run the linear programs in (3.13) and (3.14) subject to the constraint (B.13). The researcher must then repeat the procedure for all possible orderings of the support points in $\mathcal{X} \times \mathcal{Z}$. On each iteration of this procedure the researcher obtains an interval with endpoints determined by the values of the linear programs in (3.13) and (3.14). The closed convex hull of the identified set for the counterfactual probability is then given by the interval whose lower endpoint is the smallest value of the linear program in (3.13) obtained across all orderings, and whose upper endpoint is the largest value of the linear program in (3.14) obtained across all orderings. Admittedly, there are $m!$ possible orderings for $\tilde{\varphi}(X, Z, \theta)$ unless additional assumptions are imposed. This means that considering all possible orderings may be computationally burdensome.