Counterfactual Identification and Latent Space Enumeration in Discrete Outcome Models

Jiaying Gu* *University of Toronto*

Thomas M. Russell† *Carleton University*

Thomas Stringham‡ *University of Toronto*

August 11, 2022

Abstract

This paper provides a unified framework for partial identification of counterfactual parameters in a general class of discrete outcome models allowing for endogenous regressors and multidimensional latent variables, all without parametric distributional assumptions. Our main theoretical result is that, when the covariates are discrete, the infinite-dimensional latent variable distribution can be replaced with a finite-dimensional version that is equivalent from an identification perspective. The finite-dimensional latent variable distribution is constructed in practice by enumerating regions of the latent variable space with a new and efficient cell enumeration algorithm for hyperplane arrangements. We then show that bounds on a certain class of counterfactual parameters can be computed by solving a sequence of linear programming problems, and show how the researcher can introduce additional assumptions as constraints in the linear programs. Finally, we apply the method to a mobile phone choice example with heterogeneous choice sets, as well as an airline entry game example.

Keywords: Counterfactuals, Hyperplane Arrangement, Linear Programming, Partial Identification, Simultaneous Equation Models

We grateful to seminar audiences at Cornell University and Georgetown University, and especially to Francesca Molinari for her helpful comments. The first two authors acknowledge support from the Social Sciences and Humanities Research Council of Canada, grant number 435-2022-1016. 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.

‡Thomas Stringham, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S3G7, Canada. Email: tom.stringham@mail.utoronto.ca.

1 Introduction

Nonlinear models are natural in empirical settings when the outcome variable is discrete, but they often require strong parametric distributional assumptions. These assumptions impact estimated counterfactual effects, which often depend mechanically on both the structural parameters and the latent variable distribution. Flexibly modeling latent heterogeneity in nonlinear models is also complicated, and can often only be achieved under stringent assumptions.

This paper considers partial identification of counterfactual parameters in models with discrete outcomes while making minimal assumptions on the latent variable distribution. Our main assumption is that the model has an index structure that is linear in the latent variables, which is satisfied by many commonly used nonlinear structural models, and also allows for nonseparable heterogeneity. Our formulation of the problem easily allows for endogenous variables and the inclusion of a variety of additional assumptions. Our framework is applicable to models that are incomplete and incoherent, and so can be used to perform counterfactuals in nonlinear simultaneous equations models. Our main contribution is to present a unified and general procedure for bounding counterfactual effects, integrating an identification argument with state-of-the-art computational tools. Our final procedure can be used as a flexible method of modeling, and can also be used to explore sensitivity to parametric distributional assumptions commonly used in the literature.

The main challenge in our setting is that the unknown latent variable distribution is an infinite-dimensional object. A second major challenge is that the latent variable distribution is typically not nonparametrically identified in the class of models we consider, even under strong additional assumptions. To address the first challenge, we show that any latent variable distribution consistent with our assumptions can be rationalized by a discrete latent variable distribution with support contained strictly within a finite number of sets in the latent variable space. To address the second challenge, we profile over all such discrete latent variable distributions to construct bounds on counterfactual parameters. Our theoretical results demonstrate sharpness of the procedure, and nest a number of previous results as a special case.

Motivated by our theoretical results, we demonstrate how to partition the latent variable space into a finite number of disjoint sets sufficient to characterize the latent variable distribution, which we refer to as *latent space enumeration*. To perform latent space enumeration, we introduce a new recursive cell enumeration algorithm for hyperplane arrangements. We then construct all relevant constraints on the discretized latent variable distribution and bound counterfactual parameters by solving a sequence of linear programming problems. We also investigate the smallest collection of sets that are both necessary and sufficient for sharpness of our bounds, and extend the minimum relevant partition in Tebaldi et al. (2021) to a more general class of models. We then demonstrate how sign restrictions, independence assumptions, quantile restrictions, and equilibrium selection assumptions can be imposed by adding additional constraints to the linear programs. The formulation of the bounding problem is related to a recent literature on inference based on linear programs, and allows us to take advantage of previously developed consistency results.

At first glance our procedure appears to require solving two linear programming problems for each vector

of structural parameters in the identified set. In most of the existing literature the identified set for the structural parameters is found by evaluating a criterion function at a coarse grid over the parameter space, which necessarily leads to an inner-approximation to the identified set for the counterfactual parameter, even with the most sophisticated gridding strategies. Under some additional assumptions, we show that the identified set of structural parameters is given by a finite disjoint union of cones. Within a given cone, each vector of structural parameters delivers the same bounds on the counterfactual parameter, so that each cone can be viewed as an equivalence class. In order to construct bounds on our counterfactual parameters we show that the researcher need only solve two linear programs for a single vector of structural parameters from each of these cones. We call the vector of structural parameters selected from each cone a *representative point*, and we call the algorithm for finding representative points the *profiling algorithm*. By solving two linear programming problems at each representative point, we construct *exact* (i.e. not approximate) bounds for counterfactual parameters, and avoid problems that arise from blindly gridding over the parameter space. The final result is a computationally convenient and unifying procedure for bounding counterfactual outcomes that is especially useful for models where the identified set of structural parameters is difficult to derive analytically. Although our focus is on counterfactual parameters that depend on the latent variable distribution, the identified set of structural parameters can also be constructed using our linear programming approach, providing an optimization based approach for bounding structural parameters in a general class of models.

Finally, we illustrate the method with two applications: the first uses experimental data on mobile phone choice from Keane and Wasi (2013), and the second uses airline entry game data from Kline and Tamer (2016). We find informative bounds on marginal effects for a variety of specifications in both applications, and show how the bounds change under a variety of different assumptions. Both examples also illustrate how the latent space cells and representative points change under different assumptions, and suggest that a relatively small number of carefully selected representative points are sufficient to construct sharp bounds on counterfactual parameters.

1.1 Relevant Literature

There are now a variety of papers that attempt to relax assumptions on the distribution of latent variables and obtain either point or partial identification under other semiparametric restrictions. The origins of the literature in econometrics go at least as far back as Manski (1975), who introduced the maximum score estimator allowing for identification of the structural parameters in a discrete choice model without parametric distributional assumptions. The majority of papers in the literature on partial identification of structural models are concerned with the identified set of structural parameters instead of identified sets for the counterfactual parameters. The distinction is important, since in nonlinear models knowledge of the identified set of structural parameters is not enough to bound counterfactual or marginal effects without parametric distributional assumptions. Structural parameters in discrete outcome models can also

be difficult to interpret, since identification of these parameters often depends on a number of user-specified normalizations.

General approaches to bounding either structural parameters or counterfactual effects include Ekeland et al. (2010), Beresteanu et al. (2011), Galichon and Henry (2011), Schennach (2014), Chesher and Rosen (2017), Lafférs (2019), Torgovitsky (2019b), and Christensen and Connault (2019). Of these approaches, ours is most similar to the *partial identification by extending subdistributions* (PIES) approach of Torgovitsky (2019b), who also relies on a similar discretization of the latent variable space and presents optimization-based bounds on counterfactual outcomes. Our results generalize the results of Torgovitsky (2019b) in a number of ways, including allowing for incomplete models and nonseparable latent variables, which cannot be accommodated by the existing PIES framework. Furthermore, our work on latent space enumeration and the profiling algorithm for structural parameters goes far beyond the computational considerations in Torgovitsky (2019b) and casts light on a deep hyperplane structure common to a number of models in econometrics.

A number of other papers have used a discretization of the latent variable space to obtain a tractable characterization of the identified set. Manski (2007) considered partial identification of counterfactual choice probabilities in a model with linear (in attributes) utility, and possibly multidimensional latent variables. The analysis of Manski (2007) also leads to a partition of the latent variable space, although Manski (2007) does not consider implementation or computational issues, which we demonstrate are highly nontrivial. Recently Tebaldi et al. (2021) consider counterfactuals in a model of demand for California health insurance with quasi-linear indirect utility. Their model is a special case of our general framework, and many of their counterfactual parameters can be efficiently bounded using the approach in this paper. We also introduce a generalization of the *minimal relevant partition* from Tebaldi et al. (2021) that is applicable to our larger class of models and discuss how it can be constructed using our framework. Tebaldi et al. (2021) point out the need for more efficient computational tools, which is where we make one of our main contributions.[1]

Chesher et al. (2013) consider partial identification in a general class of multinomial choice models with endogenous regressors. In addition, Chesher and Rosen (2014) consider partial identification in binary response models with random coefficients, and also discover a related partition of the latent variable space. Finally, Chesher and Rosen (2020) consider interdependent discrete choice models, allowing for incompleteness and incoherency. They specifically discuss how to relax parametric distributional assumptions in these models, although they focus on structural parameters and propose outer (i.e. not sharp) bounds. The approach used in these papers is a special case of a more general framework developed in Chesher and Rosen (2017) based on a result in random set theory from Artstein (1983). The result implies a set of inequality restrictions, *Artstein's inequalities*, that must be satisfied at each vector of structural parameters in the iden-

[1] In Tebaldi et al. (2021) Section S2 when describing their approach to profiling structural parameters: "Developing a computational strategy that is feasible at scale is more challenging, but not impossible...we leave fuller investigations of these extensions to future work." Furthermore, in Section S3 when describing their approach to constructing the minimal relevant partition: "We expect that [our algorithm] leaves room for significant computational improvements, but we leave more sophisticated developments for future work."

tified set. The approach of Chesher and Rosen (2017) does not require parametric distributional assumptions on the latent variables, although in practice the number of inequalities needed to characterize the identified set can be prohibitive. Some tractable results for this case are provided by Corollary 3 in Chesher and Rosen (2017) and Theorem 2 in Chesher and Rosen (2020), although both of these results only produce outer sets. Tebaldi et al. (2021) give a detailed comparison between approaches based on the minimal relevant partition (similar to our approach) and Artstein's inequalities, and show that the minimal relevant partition can offer significant computational advantages and can deliver identified sets which are substantially smaller than the outer sets proposed by Chesher and Rosen (2017) and Chesher and Rosen (2020).

We show that the number of constraints implied by our approach is bounded by a polynomial in the support of the observed random variables, with the degree of the polynomial determined by the dimension of the latent variable space. Furthermore, Artstein's inequalities require a blind grid over the parameter space to construct the identified set for structural parameters, something we show can be avoided in certain settings.[2] Finally, and perhaps most importantly, papers using Artstein's inequalities have dealt almost exclusively with the identification of structural parameters, and not the identification of counterfactual parameters that depend on the unknown latent variable distribution. Focusing on counterfactual parameters changes many aspects of the problem: both the test sets needed for Artstein's inequalities as well as the minimal relevant partition of the latent variable space can depend on the researcher's counterfactual parameter of interest. Our approach emerges as a natural contender in the more complex environment.

The paper closest to the current paper is Gu and Russell (2021), who study partial identification in nonseparable binary response models without restricting the distribution of the latent variables while allowing for endogenous regressors. Both the theoretical and computational results in the current paper build on the insights of Gu and Russell (2021), although there are a number of substantial differences. First, the results in the current paper are applicable to a general class of nonlinear simultaneous equation models that are both *incomplete* and *incoherent*.[3] In our more general environment, some of the key insights of Gu and Russell (2021) break down, requiring a rethinking of the theoretical framework.[4] The current paper also significantly improves the computational tools used in Gu and Russell (2021) and first developed in Gu and Koenker (2020). In particular, we develop a new recursive algorithm for latent space enumeration and a new method for profiling the structural parameters. These algorithms offer vast improvements over existing methods and may be of independent interest to researchers in applied mathematics or computer science. For instance, Kitamura and Stoye (2018) develop a test of stochastic rationalizability, and constructing the rationalizable choice sets can be seen as a latent space enumeration problem. In this setting, our enumeration algorithm offers considerable computational advantages. Our approach may also be useful for implementing

[2]This is not unique to approaches based on Artstein's inequalities, as most procedures in the literature on partial identification require a blind grid to locate and characterize the identified of structural parameters.

[3]Many of the issues pertinent to our study of nonlinear simultaneous equations models—namely *incompleteness* and *incoherency*—have been discussed in Chesher and Rosen (2020) in a similar context. See also Lewbel (2019).

[4]In particular, all of the analysis in Gu and Russell (2021) is focused around a primitive counterfactual object called a *response type* (e.g. Balke and Pearl (1994), Heckman and Pinto (2018)). Model incompleteness substantially weakens the connection.

the method of Kitamura and Stoye (2019), who study nonparametric bounds on counterfactual demand under a stochastic rationalizability constraint.[5]

Our latent space enumeration algorithm connects to the literature on cell enumeration for hyperplane arrangements. Two well-known families of cell enumeration algorithms are the reverse search approach, introduced by Avis and Fukuda (1996) and further developed by Sleumer (1998), and the incremental approach, including Rada and Černý (2018) and Gu and Koenker (2020). Reverse search approaches have high computational time requirements compared to other algorithms, but are less memory-intensive. Here we use an incremental approach, which adds hyperplanes successively and records the newly-created cells at each iteration. Both incremental algorithms in Rada and Černý (2018) and Gu and Koenker (2020) use linear programs to find newly-created cells when a new hyperplane is added to an existing arrangement. In this paper we propose a new algorithm with two key innovations: (i) it uses a recursive property of hyperplane arrangements to efficiently locate newly crossed cells, and (ii) it does not rely on linear programs to identify newly-created cells at each iteration. In simulations, we show that our approach dominates current state-of-the-art algorithms by a significant margin.

1.2 Outline of Paper

The remainder of the paper proceeds as follows. Section 2 introduces the main assumptions, motivating examples, the main identification result, and our optimization-based formulation of the counterfactual bounds. Section 3 focuses on computational considerations and introduces our new enumeration and profiling algorithms. A discussion of the minimal relevant partition is included at the end of Section 3. Section 4 discusses additional assumptions that can be incorporated by our framework, including independence assumptions, quantile and sign restrictions, and counterfactual equilibrium selection assumptions. Section 5 then presents two applications of our procedure: the first application is to a mobile phone choice example using data from Keane and Wasi (2013), and the second application is to an airline entry game example using data from Kline and Tamer (2016). Section 6 concludes. The proofs of the results in the main text are in Appendix A, and some additional results and proofs are presented in Appendix B. Appendix C presents a simplified discussion of our main results when the econometric model is complete, and Appendix D discusses details related to the construction of the minimal relevant partition. Finally, Appendix E reviews a consistency result that can be used in our general setting.

[5] We thank Jörg Stoye for pointing this out.

2 Methodology

2.1 The Main Assumptions and Examples

We start by introducing our main assumptions. We use $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{U}$ to denote the random vectors:

$$\boldsymbol{Y} := \left(Y_1, Y_2, \ldots, Y_{d_y}\right)^\top, \qquad \boldsymbol{X} := \left(X_1, X_2, \ldots, X_{d_x}\right)^\top, \qquad \boldsymbol{U} := \left(U_1, U_2, \ldots, U_{d_u}\right)^\top.$$

We refer to the random variables in $\boldsymbol{Y}$ as *outcome variables*, the random variables in $\boldsymbol{X}$ as *covariates*, and the random variables in $\boldsymbol{U}$ as *latent variables*. The following assumption describes the class of models to which our method is applicable.

Assumption 2.1. *Let $\mathcal{A}_k^{(m_1,m_2)}$ denote a finite collection of finite subsets of $\mathbb{N}$ for $k = 1, \ldots, d_y$, $m_1 \in \mathcal{M}_1$, and $m_2 \in \mathcal{M}_2$, where $\mathcal{M}_1, \mathcal{M}_2 \subset \mathbb{N}$ are finite. Then there exists a complete probability space $(\Omega, \mathfrak{A}, P)$, and random vectors $\boldsymbol{Y} : \Omega \to \mathcal{Y} \subset \mathbb{R}^{d_y}$, $\boldsymbol{X} : \Omega \to \mathcal{X} \subset \mathbb{R}^{d_x}$, and $\boldsymbol{U} : \Omega \to \mathcal{U} = \mathbb{R}^{d_u}$ satisfying:*

$$\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{X}, \boldsymbol{U}, \theta_0) := \bigcap_{m_1 \in \mathcal{M}_1} \bigcup_{m_2 \in \mathcal{M}_2} \mathcal{Y}^{(m_1,m_2)}(\boldsymbol{X}, \boldsymbol{U}, \theta_0), \tag{2.1}$$

almost surely, where:

$$\mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x}, \boldsymbol{u}, \theta_0) = \left\{ \boldsymbol{y} \in \overline{\mathcal{Y}}_{(m_1,m_2)} : \boldsymbol{y} = \varphi^{(m_1,m_2)}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta_0) \right\}, \tag{2.2}$$

where $\overline{\mathcal{Y}}_{(m_1,m_2)} \subseteq \mathcal{Y}$, and where $\varphi^{(m_1,m_2)}(\,\cdot\,, \theta_0) : \mathcal{Y} \times \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$ is a $d_y \times 1$ vector-valued function with k^{th} element:

$$\varphi_k^{(m_1,m_2)}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{U}, \theta) := \sum_{A \in \mathcal{A}_k^{(m_1,m_2)}} c_k^{(m_1,m_2)}(A) \prod_{j \in A} \mathbb{1}\{g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{U}, \theta) \geq 0\}, \tag{2.3}$$

where the $c_k^{(m_1,m_2)}(A)$ are known coefficients, and where:

$$g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) := \boldsymbol{a}_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \theta)^\top \boldsymbol{u} + b_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \theta), \tag{2.4}$$

where $b_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and each element of $\boldsymbol{a}_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \theta)$ are known functions parameterized by $\theta \in \Theta = \mathbb{R}^{d_\theta}$, measurable in $(\boldsymbol{y}, \boldsymbol{x})$ for each $(\boldsymbol{u}, \theta)$. Furthermore, the spaces $\mathcal{Y}$ and $\mathcal{X}$ are finite, and $\boldsymbol{U}$ has a distribution that assigns zero probability to all sets of the form $\{\boldsymbol{u} \in \mathcal{U} : g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta_0) = 0\}$.

Following Chesher and Rosen (2017), we call the set in (2.1) the $\boldsymbol{Y}-$level set.[6] The set contains all possible values of the outcome variables consistent with a tuple $(\boldsymbol{x}, \boldsymbol{u}, \theta)$ under Assumption 2.1. Intuitively, when $(\boldsymbol{x}, \boldsymbol{u})$ are replaced with the random vectors $(\boldsymbol{X}, \boldsymbol{U})$ the $\boldsymbol{Y}-$level set in (2.1) becomes a *random set*.[7] A measurable selection from the random set (2.1) is a random vector $\boldsymbol{Y}$ satisfying $\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{X}, \boldsymbol{U}, \theta)$ almost surely, and any selection of the $\boldsymbol{Y}-$level set can be viewed as a possible outcome of the econometric

[6] Other names have been provided; for instance, Jovanovic (1989) calls it the "reduced-form correspondence" or the "equilibrium correspondence."

[7] A random set is formally defined as an Effros-measurable multifunction on a probability space (see Definition B.2). Lemma B.1 verifies Effros-measurability of the $\boldsymbol{Y}-$level set under Assumption 2.1.

model. If the econometric model is incomplete, then the $\boldsymbol{Y}-$level sets can contain multiple elements. If the model is incoherent, then the $\boldsymbol{Y}-$level set can be empty. In a general system of nonlinear equations both incompleteness and incoherency must be admitted as possibilities.[8] Assumption 2.1 allows the model to be both *incomplete* and *incoherent*.

Assumption 2.1 restricts attention to models where the outcomes and covariates are discrete, and where the outcome variables can be determined by the nonlinear simultaneous equations model in (2.2) where each function in the system is of the form (2.3).[9] Assumption 2.1 is satisfied by a general class of nonlinear simultaneous equations models, covering a number of commonly used models as a special case. The inclusion of the index sets $\mathcal{M}_1$ and $\mathcal{M}_2$ allow us to construct more sophisticated $\boldsymbol{Y}-$level sets, expanding the set of applicable models. We illustrate the use of these index sets in the examples ahead. In equation (2.4), we impose that each index function from (2.3) is linear in the vector $\boldsymbol{u}$ representing the latent variables. This assumption is key to our approach, but it accommodates most commonly used specifications in the models we have in mind, including models that allow for random coefficients. The restriction to covariates with finite support is necessary for computational reasons, but it also simplifies the theoretical results considerably. The assumption is also maintained in nearly all of the related papers in the partial identification literature, either for theoretical or computational reasons, and we do not innovate on the existing literature in this regard.[10] The final part of the assumption rules out cases when the functions $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ from (2.4) are identically zero with positive probability. This assumption is implied by absolute continuity of the distribution of latent variables with respect to the Lebesgue measure—a standard assumption in this literature—and still allows the latent variable distribution to contain mass points.

Example 1 (Multinomial Choice/Demand)**.** *Consider the following example of multinomial choice. Consumer i faces a set of alternatives $\mathcal{J}_i \subseteq \{1,2,\ldots,J\}$, and chooses the alternative that maximizes their utility. Denote the consumer's observed choice by Y_i, suppose that $|\mathcal{J}_i| \geq \kappa$ for some known integer $\kappa \geq 1$, and let $\mathscr{J}_\kappa$ denote the collection of all subsets of $\{1,2,\ldots,J\}$ of size at least κ, with typical element $\mathcal{J} \in \mathscr{J}_\kappa$. The econometrician knows that consumers face a choice set in $\mathscr{J}_\kappa$, but does not observe the specific choice set faced by each consumer. Let ζ_{ij} denote the utility obtained when consumer i chooses alternative j. A typical specification for the consumer utility function is:*

$$\zeta_{ij} = \boldsymbol{X}_{ij}^\top \theta + U_{ij},$$

where $\boldsymbol{X}_{ij}$ is a vector of attributes for product j that may also vary by individual, θ is a vector of structural parameters shared by all individuals, and U_{ij} is an individual-product-specific utility shock. Given the choice

[8] For comparable definitions, see Tamer (2003) and Lewbel (2019). Also, see Chesher and Rosen (2020) for a recent and relevant discussion of these issues.

[9] The formulation in (2.3) was motivated by the model in equation (2.2) of Lee and Salanié (2018), which they show can accommodate a variety of discrete outcome models.

[10] For instance, in the relevant papers on binary choice (e.g. Chesher and Rosen (2014), Gu and Russell (2021)), papers on multinomial choice (e.g. Chesher et al. (2013), Tebaldi et al. (2021)), papers on dynamic binary response models (e.g. Honoré and Tamer (2006), Chernozhukov et al. (2013), Torgovitsky (2019a)), and papers on games (e.g. Ciliberto and Tamer (2009), Kline and Tamer (2016)), as well as more general methods (e.g. Chesher and Rosen (2017), Torgovitsky (2019b)).

Example	d_y	$\mathcal{M}_1$	$\mathcal{M}_2$	$\overline{\mathcal{Y}}_{(m_1,m_2)}$	$\mathcal{A}_k^{(m_1,m_2)}$	$c_k^{(m_1,m_2)}(A)$	$g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$
Multinomial Choice	1	$\varnothing$	Index set for $\mathcal{J}\subset\{1,\ldots,J\}$ with $\lvert\mathcal{J}\rvert\geq\kappa$.	$\{1,2,\ldots,J\}$	$\bigcup_{j'}\{\mathcal{J}\setminus\{j'\}\}$	$c_k^{(m_1,m_2)}(\mathcal{J}\setminus\{j'\})=j'$	$(\boldsymbol{x}_{j'}-\boldsymbol{x}_j)^\top\theta+u_{j'}-u_j$
Game (Nash Equilibria)	K	$\{1\}$	$\{1\}$	$\{0,1\}^K$	$\{\{1\}\}$	$c_k^{(m_1,m_2)}(\{1\})=1$	$\boldsymbol{x}_k^\top\beta_k+\boldsymbol{y}_{(-k)}^\top\delta_k-u_k$
Game (Level-1 Rationality)	K	$\{1,2,\ldots,K\}$	Index set for elements of $\{0,1\}^{K-1}$.	$\{0,1\}^K$	$\{\{1\}\}$	$c_k^{(m_1,m_2)}(\{1\})=1$	$\boldsymbol{x}_k^\top\beta_k+\boldsymbol{y}_{(-k)}^\top\delta_k-u_k$
Dynamic Binary Response	T	$\{1\}$	$\{1\}$	$\{0,1\}^T$	$\{\{1\}\}$	$c_k^{(m_1,m_2)}(\{1\})=1$	$\boldsymbol{x}_k^\top\beta+y_{k-1}\delta+u_0+u_k$

Table 1: Specification of the main components of the general model in Assumption 2.1 for examples 1, 2 and 3.

set $\mathcal{J}\in\mathscr{J}_\kappa$, consumer i chooses the alternative to maximize utility:

$$Y_i=\sum_{j'\in\mathcal{J}}j'\cdot\prod_{j\in\mathcal{J}\setminus\{j'\}}\mathbb{1}\{\boldsymbol{X}_{ij'}^\top\theta+U_{ij'}\geq\boldsymbol{X}_{ij}^\top\theta+U_{ij}\}. \tag{2.5}$$

Let $\mathcal{Y}^{\mathcal{J}}(\boldsymbol{x},\boldsymbol{u},\theta)$ denote all values of $\boldsymbol{y}\in\mathcal{Y}=\{1,2,\ldots,J\}$ satisfying (2.5) for a given tuple $(\boldsymbol{x},\boldsymbol{u},\theta)$, where $\boldsymbol{x}=(\boldsymbol{x}_j)_{j=1}^J$ and $\boldsymbol{u}=(u_j)_{j=1}^J$ are both $J\times 1$ vectors. Then (2.5) restricts $Y_i=\mathcal{Y}^{\mathcal{J}}(\boldsymbol{x},\boldsymbol{u},\theta)$, a singleton set. When $\kappa=J$, all consumers face the same (full) choice set. When $\kappa<J$ consumers face a choice set $\mathcal{J}_i$ belonging to the collection $\mathscr{J}_\kappa$. Define:

$$\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta):=\bigcup_{\mathcal{J}\in\mathscr{J}_\kappa}\mathcal{Y}^{\mathcal{J}}(\boldsymbol{x},\boldsymbol{u},\theta)=\bigcup_{\mathcal{J}\subset\{1,\ldots,J\}:\lvert\mathcal{J}\rvert\geq\kappa}\mathcal{Y}^{\mathcal{J}}(\boldsymbol{x},\boldsymbol{u},\theta). \tag{2.6}$$

Then the model restricts $Y_i\in\mathcal{Y}(\boldsymbol{X}_i,\boldsymbol{U}_i,\theta_0)$, meaning the observed choice Y_i was optimal given some latent choice set $\mathcal{J}$ of size at least $\kappa\geq 1$. Table 1 specifies all the components of Assumption 2.1 required to fit the example.

Typical multinomial choice models (c.f. McFadden (1974)) specify an additively separable error term with a parametric distribution which is known up to a finite-dimensional parameter. In contrast, our framework allows for multidimensional latent variables without making parametric distributional assumptions, while allowing for various semiparametric restrictions. Our paper also easily accommodates endogenous covariates and random coefficients. Relevant papers that also focus on relaxing distributional assumptions include Manski (1975), Fox (2007), Manski (2007), Chesher et al. (2013), Chesher and Rosen (2014), Pakes et al. (2015), Chiong et al. (2017), Allen and Rehbeck (2019), and Tebaldi et al. (2021). The framework in (2.6) for studying unobserved and heterogeneous choice sets was proposed by Barseghyan et al. (2021), with a similar setting also studied by Lu (2022). These papers focus on constructing the identified set for structural parameters when the distribution of latent variables is parametrically specified. In this paper, we show how to bound counterfactual parameters in this model when the distribution of latent variables is not specified.

Example 2 (Static Games of Complete Information)**.** *Consider a static game of complete information where*

the action space is binary. We consider only pure strategy equilibria. Suppose there are K players whose payoffs in market i are determined as follows:

$$\pi_{ik} = \begin{cases} \boldsymbol{X}_{ik}^\top \beta_k + \boldsymbol{Y}_{i(-k)}^\top \delta_k - U_{ik}, & \text{if } Y_{ik} = 1, \\ 0, & \text{if } Y_{ik} = 0, \end{cases}$$

where $\boldsymbol{Y}_{i(-k)}$ is a $(K-1) \times 1$ binary vector representing the actions of all players other than player k, and where $\boldsymbol{U}_i = (U_{ik})_{k=1}^K$ is a $K \times 1$ vector of payoff shocks. In a complete information setting the payoff functions are observed by all players. Player k chooses the action that maximizes their payoff, but their payoff-maximizing choice depends on the actions chosen by the other players. A best response for player k is a strategy that dominates all other strategies, given a strategy profile for the other players. For a given vector of structural parameters, and a given realization $\boldsymbol{x} = (\boldsymbol{x}_k)_{k=1}^K$ of the covariates and $\boldsymbol{u} = (u_k)_{k=1}^K$ of the latent variables, the set of Nash equilibria to the game belong to the following set:

$$\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \left\{ \boldsymbol{y} \in \{0,1\}^K : \begin{array}{c} y_1 = \mathbb{1}\{\boldsymbol{x}_1^\top \beta_1 + \boldsymbol{y}_{(-1)}^\top \delta_1 - u_1 \geq 0\} \\ y_2 = \mathbb{1}\{\boldsymbol{x}_2^\top \beta_2 + \boldsymbol{y}_{(-2)}^\top \delta_2 - u_2 \geq 0\} \\ \vdots \\ y_K = \mathbb{1}\{\boldsymbol{x}_K^\top \beta_K + \boldsymbol{y}_{(-K)}^\top \delta_K - u_K \geq 0\} \end{array} \right\}.$$

Table 1 shows how to specify the components of Assumption 2.1 to fit the example. Our framework also allows for alternative solution concepts. For instance, following Bernheim (1984) and Pearce (1984), a level-1 rational strategy for player k is a best response to some strategy profile of the other players. A level-s rational strategy for player k is a best response to some level-$(s-1)$ rational strategy profile of the other players.[11] *For a given vector of structural parameters, and a given realization of the covariates and latent variables the level-1 rational outcomes of the game are contained in the set:*[12]

$$\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta_0) := \bigcap_{k=1}^K \bigcup_{\tilde{\boldsymbol{y}} \in \{0,1\}^{K-1}} \left\{ \boldsymbol{y} \in \{0,1\}^K : y_k = \mathbb{1}\{\boldsymbol{x}_k^\top \beta_k + \tilde{\boldsymbol{y}}^\top \delta_k - u_k \geq 0\} \right\}.$$

Table 1 also shows how to specify the components of Assumption 2.1 to fit the example with level-1 rational outcomes. Level-s rational outcomes can also be accommodated by a slight extension of Assumption 2.1.

The literature on these models is now quite large, with relevant contributions made by Bjorn and Vuong (1984), Bresnahan and Reiss (1991a,b), Berry (1992), Tamer (2003), Molinari and Rosen (2008), Ciliberto

[11] Here we focus on pure strategies, although the rationality concept of Bernheim (1984) and Pearce (1984) allows for mixed strategies.

[12] For instance, suppose that $K = 2$, $\delta_1, \delta_2 < 0$, $u_1 \leq \boldsymbol{x}_1^\top \beta_1 + \delta_1$ and $\boldsymbol{x}_2^\top \beta_2 + \delta_2 \leq u_2 \leq \boldsymbol{x}_2^\top \beta_2$. Then $y_1 = 1$ is a dominant strategy for player 1. Furthermore, $y_2 = 1$ is a best response to $y_1 = 0$, and $y_2 = 0$ is a best response to $y_1 = 1$. We have:

$$\begin{aligned} \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta_0) &= \bigcap_{k=1}^2 \bigcup_{\tilde{\boldsymbol{y}} \in \{0,1\}} \left\{ \boldsymbol{y} \in \{0,1\}^2 : y_k = \mathbb{1}\{\boldsymbol{x}_k^\top \beta_k + \tilde{y}^\top \delta_k - u_k \geq 0\} \right\} \\ &= (\{(1,0),(1,1)\} \cup \{(1,0),(1,1)\}) \cap (\{(0,1),(1,1)\} \cup \{(0,0),(1,0)\}) \\ &= \{(1,0),(1,1)\}. \end{aligned}$$

and Tamer (2009), Bajari et al. (2010), Beresteanu et al. (2011), Galichon and Henry (2011), Dunker et al. (2018), Jun and Pinkse (2020), among many others. We refer the reader to Aradillas-López (2020) for a detailed and recent review of the econometric issues in these models. Level-s rationality has been explored in empirical games by Aradillas-López and Tamer (2008) and Kline and Tamer (2012).

Example 3 (Dynamic Panel Data Discrete Response Model)**.** *Consider the following discrete response panel data model with time-invariant heterogeneity:*

$$\begin{aligned} Y_{i1} &= \mathbb{1}\{\boldsymbol{X}_{i1}^\top \beta + Y_{i0}\delta + U_{i0} + U_{i1} \geq 0\}, \\ Y_{i2} &= \mathbb{1}\{\boldsymbol{X}_{i2}^\top \beta + Y_{i1}\delta + U_{i0} + U_{i2} \geq 0\}, \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_{iT} &= \mathbb{1}\{\boldsymbol{X}_{iT}^\top \beta + Y_{iT-1}\delta + U_{i0} + U_{iT} \geq 0\}, \end{aligned}$$

where $\{\boldsymbol{X}_{it}\}_{t=1}^T$ is a sequence of observed vectors of covariates, $\{Y_{it}\}_{t=1}^T$ is a sequence of observed binary outcomes, Y_{i0} represents the period zero outcome, $\{U_{it}\}_{t=1}^T$ is a sequence of latent and transitory random shocks, and U_{i0} is a random variable representing time-invariant unobserved heterogeneity. See Honoré and Kyriazidou (2000) for a similar environment. If we treat the initial conditions Y_{i0} as unobserved the model is incomplete (c.f. Honoré and Tamer (2006)). To accommodate this model, for a given vector of structural parameters, and a given vector of covariates $\boldsymbol{x} = (\boldsymbol{x}_t)_{t=1}^T$ and latent variables $\boldsymbol{u} = (\boldsymbol{u}_t)_{t=0}^T$, define the sets:

$$\mathcal{Y}^{(y)}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \left\{ \boldsymbol{y} \in \{0,1\}^T : \begin{array}{rcl} y_1 &=& \mathbb{1}\{\boldsymbol{x}_1^\top \beta + y\delta + u_0 + u_1 \geq 0\}, \\ y_2 &=& \mathbb{1}\{\boldsymbol{x}_2^\top \beta + y_1\delta + u_0 + u_2 \geq 0\}, \\ \vdots & & \vdots \\ y_T &=& \mathbb{1}\{\boldsymbol{x}_T^\top \beta + y_{T-1}\delta + u_0 + u_T \geq 0\}. \end{array} \right\}, \qquad \textit{(i.e. } y_0 = y\textit{)},$$

and let $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}^{(0)}(\boldsymbol{x}, \boldsymbol{u}, \theta) \cup \mathcal{Y}^{(1)}(\boldsymbol{x}, \boldsymbol{u}, \theta)$. Table 1 shows how to specify the remaining components of Assumption 2.1.

Papers studying partial identification of the dynamic binary response model include Honoré and Tamer (2006), Chernozhukov et al. (2013), Torgovitsky (2019a), Aguirregabiria and Carro (2021), Aristodemou (2021), Dobronyi et al. (2021), and Khan et al. (2021). We refer to Honoré and de Paula (2021) for a recent review.

2.2 Identification

The $\boldsymbol{Y}-$level sets are useful for defining the identified set of *structures*, which in our case are admissible pairs $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta)$ of conditional latent variable distributions $P_{\boldsymbol{U}|\boldsymbol{X}}$ and structural parameters θ.[13]

[13] We also refer to Jovanovic (1989) for additional motivation for the analysis presented here, as well as Chesher and Rosen (2017) for the notion of structures and generalized notion of identification of structures in incomplete and partially identified models.

Definition 2.1 (Identified Set)**.** *The (joint) identified set $\mathcal{I}^*_{\boldsymbol{X}}$ of structures is the set of all pairs $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta)$ such that there exists a conditional distribution $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ satisfying:*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_{B} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{2.7}$$

$P_{\boldsymbol{X}}-$almost surely for every $B \subset \mathcal{Y}$, and such that $P_{\boldsymbol{U}} = P_{\boldsymbol{U}|\boldsymbol{X}} P_{\boldsymbol{X}}$ assigns zero probability to all sets of the form $\{\boldsymbol{u} \in \mathcal{U} : g^{(m_1,m_2)}_{j,k}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\}$ from Assumption 2.1.

The definition is nearly identical to Definition 3 in Galichon and Henry (2011), although those authors dealt with the case when the latent variable distribution was parametrically specified. When the model is incomplete, $P_{Y|\boldsymbol{X},\boldsymbol{U}}$ represents the equilibrium selection mechanism. Note the identified set implicitly depends on the distribution of $\boldsymbol{X}$ through the almost-sure relation in (2.7); any values of $\boldsymbol{x}$ assigned zero probability by the observed distribution do not impose restrictions on the distribution of $\boldsymbol{U}$. Although we are not focused on the identification of structures, the definition will be useful to motivate the results ahead.

In our setting, each counterfactual is represented by a tuple:

$$\gamma := \left(c_x, \left(\boldsymbol{y}^\star_{(m_1,m_2)} \right)_{m_1 \in \mathcal{M}_1, m_2 \in \mathcal{M}_2}, \left(\Pi^\star_{(m_1,m_2)} \right)_{m_1 \in \mathcal{M}_1, m_2 \in \mathcal{M}_2}, \left(\overline{\mathcal{Y}}^\star_{(m_1,m_2)} \right)_{m_1 \in \mathcal{M}_1, m_2 \in \mathcal{M}_2}, \mathcal{M}^\star_1, \mathcal{M}^\star_2 \right),$$

where $c_x : \mathcal{X} \to \mathcal{X}$ is a function, and $\boldsymbol{y}^\star_{(m_1,m_2)} \in \mathcal{Y}$ is a vector, $\Pi^\star_{(m_1,m_2)} \subseteq \{1, \ldots, d_y\}$ is an index set, $\overline{\mathcal{Y}}^\star_{(m_1,m_2)} \subseteq \mathcal{Y}$ is a subset, and $\mathcal{M}^\star_1 \subseteq \mathcal{M}_1$ and $\mathcal{M}^\star_2 \subseteq \mathcal{M}_2$ are index sets. The elements of the tuple γ represent modifications of the econometric model specified in Assumption 2.1.

Assumption 2.2. *For the counterfactual represented by γ, there exists a random vector $\boldsymbol{Y}_\gamma : \Omega \to \mathcal{Y}$ (the counterfactual outcome vector) satisfying:*

$$\boldsymbol{Y}_\gamma \in \mathcal{Y}_\gamma(\boldsymbol{X}, \boldsymbol{U}, \theta_0) := \bigcap_{m_1 \in \mathcal{M}^\star_1} \bigcup_{m_2 \in \mathcal{M}^\star_2} \mathcal{Y}^{(m_1,m_2)}_\gamma(\boldsymbol{X}, \boldsymbol{U}, \theta_0), \tag{2.8}$$

for the same $\theta_0 \in \Theta$ as in Assumption 2.1, where $\mathcal{M}^\star_1 \subseteq \mathcal{M}_1$, $\mathcal{M}^\star_2 \subseteq \mathcal{M}_2$, and:

$$\mathcal{Y}^{(m_1,m_2)}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) = \left\{ \boldsymbol{y} \in \overline{\mathcal{Y}}^\star_{(m_1,m_2)} : \boldsymbol{y} = \varphi^{(m_1,m_2)}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \right\},$$

where $\overline{\mathcal{Y}}^\star_{m_1,m_2} \subseteq \mathcal{Y}$, and where $\varphi^{(m_1,m_2)}_\gamma(\,\cdot\,, \theta) : \mathcal{Y} \times \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$ is a $d_y \times 1$ vector-valued function with k^{th} element:

$$[\varphi^{(m_1,m_2)}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)]_k = \begin{cases} [\varphi(\boldsymbol{y}, c_x(\boldsymbol{x}), \boldsymbol{u}, \theta)]_k, & \text{if } k \notin \Pi^\star_{(m_1,m_2)}, \\ [\boldsymbol{y}^\star_{(m_1,m_2)}]_k, & \text{if } k \in \Pi^\star_{(m_1,m_2)}, \end{cases}$$

for each $m_1 \in \mathcal{M}^\star_1$ and $m_2 \in \mathcal{M}^\star_2$.

Assumption 2.2 posits a collection of counterfactual random vectors $\boldsymbol{Y}_\gamma$ that are required to satisfy certain counterfactual support conditions. The support conditions determine the values of the outcome variable that satisfy the system of structural equations in (2.2) after a modification to the system has occurred. The set (2.8) is the counterfactual analog of the $\boldsymbol{Y}-$level sets from (2.1). It specifies the values of the outcome

variable consistent with the model defined in Assumption 2.1 and the counterfactual represented by the tuple γ. The tuple contains the information required to execute a counterfactual. The function c_x is used to replace the values of $\boldsymbol{X}$ with counterfactual values. The elements $\boldsymbol{y}^\star_{(m_1,m_2)}$ and $\Pi^\star_{(m_1,m_2)}$ are used to edit the structural equations to perform counterfactuals that modify the endogenous outcome variables: the index set $\Pi^\star_{(m_1,m_2)}$ determines which of the equations in the system $\boldsymbol{y} = \varphi_\gamma^{(m_1,m_2)}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta_0)$ have been "wiped out" (e.g. Strotz and Wold (1960)) and replaced with an appropriate entry of the vector $\boldsymbol{y}^\star_{(m_1,m_2)}$ specified in the counterfactual.[14] These counterfactuals are needed, for example, to study best response functions in a pure strategy game of complete information (c.f. Kline and Tamer (2012)). The subsets $\mathcal{Y}^\star_{(m_1,m_2)} \subseteq \mathcal{Y}$ are modifications of $\mathcal{Y}_{(m_1,m_2)} \subseteq \mathcal{Y}$ from Assumption 2.1 which can be used to restrict the support of the counterfactual outcome variable. Finally, the index sets $\mathcal{M}_1^\star, \mathcal{M}_2^\star \subset \mathbb{N}$ are modifications of the index sets $\mathcal{M}_1, \mathcal{M}_2 \subset \mathbb{N}$ from Assumption 2.1. The modified index sets $\mathcal{M}_1^\star, \mathcal{M}_2^\star \subset \mathbb{N}$ can be used, for example, to study counterfactuals that modify choice sets.

Remark 2.1. *For simplicity, we will omit redundant elements of the tuple γ. For instance, we omit c_x if c_x is the identity function, we omit both $\boldsymbol{y}^\star_{(m_1,m_2)}$ and $\Pi^\star_{(m_1,m_2)}$ if $\Pi^\star_{(m_1,m_2)} = \varnothing$, we omit $\overline{\mathcal{Y}}^\star_{(m_1,m_2)}$ if $\overline{\mathcal{Y}}^\star_{(m_1,m_2)} = \overline{\mathcal{Y}}_{(m_1,m_2)}$, and we omit $\mathcal{M}_1^\star$ if $\mathcal{M}_1^\star = \mathcal{M}_1$ and omit $\mathcal{M}_2^\star$ if $\mathcal{M}_2^\star = \mathcal{M}_2$. As an example, if $\gamma = (c_x, \mathcal{M}_1^\star)$, then only c_x and $\mathcal{M}_1^\star$ are relevant for executing the counterfactual. The tuple γ may still contain redundant information; for instance, depending on $\Pi^\star_{(m_1,m_2)}$, some elements of $\boldsymbol{y}^\star_{(m_1,m_2)}$ may not matter for the counterfactual exercise. The elements in $\boldsymbol{y}^\star_{(m_1,m_2)}$ not indexed by $\Pi^\star_{(m_1,m_2)}$ can be set to any value. This will be illustrated in Example 2 ahead.*

Remark 2.2. *While interventions are indexed by a function $c_x : \mathcal{X} \to \mathcal{X}$, our notation does not rule out interventions that consider covariate values outside the empirical support of $\boldsymbol{X}$. These can be accommodated by expanding the theoretical support $\mathcal{X}$ to include any additional values of interest.*[15] *A similar comment applies to counterfactual values for the outcome variables.*

We use Assumption 2.2 as a starting point to construct the identified set for counterfactual quantities of the form:

$$\Psi_{\boldsymbol{x}} := \int_{\mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}, \tag{2.9}$$

where $\psi_{\boldsymbol{x}} : \mathcal{Y} \to \mathbb{R}$ is a measurable function for each $\boldsymbol{x}$. We refer to this quantity as a *counterfactual functional*, since it is a linear functional of the counterfactual conditional distribution $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. Although we focus on functionals of the form in (2.9), our approach is equally amenable to bounding any linear functions of these counterfactual functionals, as well as linear functions that include objects of the form (2.9) for different counterfactuals γ_1 and γ_2 (e.g. marginal effects).

[14] A similar role to "wiping-out" is played by Judea Pearl's do-operator in non-recursive systems (c.f. Pearl (2009)). Also, the "role indicator" in the settable systems framework of White and Chalak (2009) plays a similar role to the index set $\Pi^\star_{(m_1,m_2)}$. See also their discussion of partitioning in settable systems.

[15] Note here the different treatment between the observed or empirical support of $\boldsymbol{X}$, which includes only values that can be assigned positive probability by $P_{\boldsymbol{X}}$, and the theoretical support of $\boldsymbol{X}$, which can include counterfactual values assigned zero probability by $P_{\boldsymbol{X}}$.

Example 1 (Multinomial Choice/Demand (cont'd)). *Consider again the example of multinomial choice with heterogeneous choice sets. Given a latent choice set of size at least $\kappa \geq 1$ the choice Y_i for consumer i satisfies:*

$$Y_i \in \mathcal{Y}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta) = \bigcup_{\mathcal{J} \subset \{1,\ldots,J\} : |\mathcal{J}| \geq \kappa} \mathcal{Y}^{\mathcal{J}}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta), \tag{2.10}$$

where $\mathcal{Y}^{\mathcal{J}}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta)$ represents the choice made given choice set $\mathcal{J}$. Indexing the choice sets in $\{\mathcal{J} \subset \{1,\ldots,J\} : |\mathcal{J}| \geq \kappa\}$ as $\mathcal{J}_1, \ldots, \mathcal{J}_M$, we can rewrite (2.10) as:

$$Y_i \in \mathcal{Y}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta) = \bigcup_{m \in \mathcal{M}_2} \mathcal{Y}^{\mathcal{J}_m}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta),$$

where $\mathcal{M}_2 = \{1, \ldots, M\}$. Without loss of generality, suppose that the full choice set is given by $\mathcal{J}_M = \{1, \ldots, J\}$, and consider a counterfactual that assigns all individuals the full choice set. In this counterfactual the choice $Y_{i\gamma}$ for consumer i satisfies:

$$Y_{i\gamma} \in \mathcal{Y}_\gamma(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta) = \bigcup_{m \in \mathcal{M}_2^\star} \mathcal{Y}^{\mathcal{J}_m}(\boldsymbol{X}_i, \boldsymbol{U}_i, \theta),$$

where $\mathcal{M}_2^\star = \{M\}$. The counterfactual is accommodated by Assumption 2.2 by setting $\gamma = (\mathcal{M}_2^\star)$. A possible counterfactual functional of interest is the counterfactual conditional choice probability:

$$P_{Y_\gamma|\boldsymbol{X}}(Y_\gamma = j \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\{1,\ldots,J\}} \mathbb{1}\{Y_\gamma = j\} \, dP_{Y_\gamma|\boldsymbol{X}},$$

for various $j \in \{1, \ldots, J\}$.

Example 2 (Static Games of Complete Information (cont'd)). *Consider a static game of complete information with pure strategy Nash equilibria with two players. The choices of the two players are determined by the simultaneous discrete choice model:*

$$Y_{i1} = \mathbb{1}\{X_{i1}\beta_1 + Y_{i2}\delta_1 - U_{i1} \geq 0\},$$
$$Y_{i2} = \mathbb{1}\{X_{i2}\beta_2 + Y_{i1}\delta_2 - U_{i2} \geq 0\}.$$

Now consider a counterfactual that sets the value of X_{i2} to x (e.g. a government subsidy on the entry of the second player), but keeps all other variables unchanged. The counterfactual can be represented by $\gamma = (c_x)$ where $c_x(x_1, x_2) = (x_1, x)$ for every $(x_1, x_2) \in \mathcal{X}$. A possible counterfactual outcome of interest is the counterfactual conditional entry probability of the first player:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(Y_{1\gamma} = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{Y}} \mathbb{1}\{Y_{1\gamma} = 1\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}.$$

Alternatively, we can consider the marginal effect of player 2's entry on player 1's entry. Let $\gamma_1 = (\boldsymbol{y}_1^\star, \Pi_1^\star)$, where $\boldsymbol{y}_1^\star = (\cdot\,, 1)$, and $\Pi_1^\star = \{2\}$ (note the first entry of $\boldsymbol{y}_1^\star$ is redundant when $\Pi_1^\star = \{2\}$). Furthermore, let $\gamma_2 = (\boldsymbol{y}_2^\star, \Pi_2^\star)$, where $\boldsymbol{y}_2^\star = (\cdot\,, 0)$, and $\Pi_2^\star = \{2\}$ (again the first entry of $\boldsymbol{y}_2^\star$ is redundant when $\Pi_1^\star = \{2\}$).

Since both $\Pi_1^\star = \Pi_2^\star = \{2\}$ *are nonempty the counterfactuals* γ_1 *and* γ_2 *modify the existing system of structural equations. The modified system under both* γ_1 *and* γ_2 *is of the form:*

$$Y_{i1} = \mathbb{1}\{X_{i1}\beta_1 + Y_{i2}\delta_1 - U_{i1} \geq 0\},$$
$$Y_{i2} = y_2,$$

where $y_2 = 1$ *under counterfactual* γ_1 *and* $y_2 = 0$ *under counterfactual* γ_2*. The marginal effect of player 2's entry on player 1's entry can then be written as:*

$$P_{\boldsymbol{Y}_{\gamma_1}}(Y_{1\gamma_1} = 1) - P_{\boldsymbol{Y}_{\gamma_2}}(Y_{1\gamma_2} = 1) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}\{Y_{1\gamma_1} = 1\}\, dP_{\boldsymbol{Y}_{\gamma_1}|\boldsymbol{X}}\, dP_{\boldsymbol{X}} - \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}\{Y_{1\gamma_2} = 1\}\, dP_{\boldsymbol{Y}_{\gamma_2}|\boldsymbol{X}}\, dP_{\boldsymbol{X}},$$

which is the difference of two linear functions of the counterfactual functional (2.9)*.*

Using Assumption 2.2, we have the following definition of the identified set for counterfactual conditional distributions, and counterfactual functionals of the form (2.9).

Definition 2.2 (Identified Set of Counterfactual Conditional Distributions)**.** *Under Assumptions 2.1 and 2.2, for a given tuple* γ *the identified set of counterfactual conditional distributions* $\mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ *is the set of all conditional distributions* $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ *such that there exists a conditional distribution* $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ *and some pair* $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ *satisfying:*

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}}\int_{B} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\}\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{2.11}$$

$P_{\boldsymbol{X}}-$*almost surely for every* $B \subset \mathcal{Y}$*. Furthermore, the identified set for the counterfactual functional* $\Psi_{\boldsymbol{x}}$ *from* (2.9) *is the set of all values* $\overline{\Psi}_{\boldsymbol{x}}$ *such that there exists some* $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ *satisfying:*

$$\overline{\Psi}_{\boldsymbol{x}} := \int_{\mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}},$$

$P_{\boldsymbol{X}}-$*almost surely.*

Note the definition also makes explicit reference to the identified set in Definition 2.1, which in turn is derived from the selection relation in Assumption 2.1. Definition 2.2 also connects the identified set $\mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ to the identified set for the counterfactual functional $\Psi_{\boldsymbol{x}}$.

Definition 2.2 suggests that constructing computationally tractable and sharp bounds on the counterfactual functional from (2.9) requires a computationally tractable and sharp characterization of the identified set $\mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. This requires a tractable means of verifying a pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ and a collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ satisfy the conditions from Definition 2.2. However, both $P_{\boldsymbol{U}|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ are *infinite-dimensional* objects, meaning that verifying a pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ and a collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ satisfy the conditions in Definition 2.2 is an *infinite-dimensional existence problem*.[16]

To make progress, our first main result replaces the infinite-dimensional existence problem with an

[16] This terminology seems to have first appeared in Torgovitsky (2019b).

equivalent finite-dimensional existence problem. To introduce the result, consider the set:

$$\left\{\boldsymbol{u} \in \mathcal{U} : g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\right\}, \tag{2.12}$$

where the function $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$ is from Assumption 2.1. By (2.4) in Assumption 2.1, these sets are hyperplanes in $\mathcal{U} = \mathbb{R}^{d_u}$. Label these hyperplanes as $H_1, H_2, \ldots, H_{n_h}$. The collection $\mathscr{H} := \{H_1, H_2, \ldots, H_{n_h}\}$ is called a *hyperplane arrangement*. The complement of the arrangement is:

$$\mathcal{U} - \bigcup_{H \in \mathscr{H}} H,$$

and consists of a finite number of open, connected, and disjoint regions called *cells*. Let $\mathscr{C}$ denote the collection of all cells, enumerated as $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_L$. Each set $\mathcal{U}_\ell$ is the intersection of open half-spaces defined by the hyperplanes in $\mathscr{H}$, and for each $\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{U}_\ell$ we have $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta)$.

Example 2 (Cont'd)**.** *Consider again a game example with two players, complete information, and pure strategy Nash equilibria, but suppose now that the players' actions are determined by the following system:*

$$Y_{i1} = \mathbb{1}\{Y_{i2}U_{i3} - U_{i1} \geq 0\},$$
$$Y_{i2} = \mathbb{1}\{Y_{i1}U_{i3} - U_{i2} \geq 0\}.$$

Here there are no structural parameters, and U_{i3} represents a random strategic interaction effect common to both players. The hyperplanes are given by:

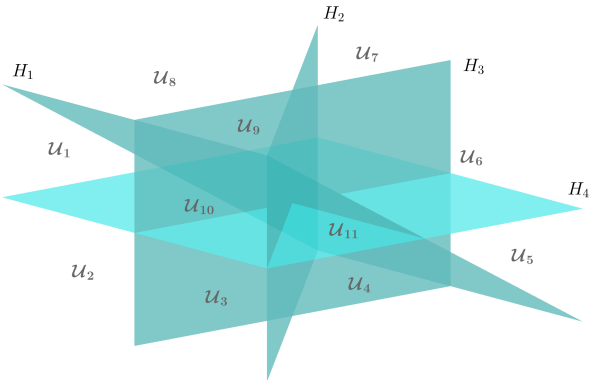
$$H_1 = \{\boldsymbol{u} \in \mathbb{R}^3 : u_3 - u_1 = 0\}, \qquad H_3 = \{\boldsymbol{u} \in \mathbb{R}^3 : u_3 - u_2 = 0\},$$
$$H_2 = \{\boldsymbol{u} \in \mathbb{R}^3 : -u_1 = 0\}, \qquad H_4 = \{\boldsymbol{u} \in \mathbb{R}^3 : -u_2 = 0\}.$$

Here $d_u = 3$ and $n_h = 4$. There are 14 cells associated with the arrangement, which are illustrated in Figure 1(a). Furthermore, all (u_1, u_2, u_3) belonging to the same cell deliver the same $\boldsymbol{Y}-$level set. The $\boldsymbol{Y}-$level sets associated with each cell are displayed in Figure 1(b).

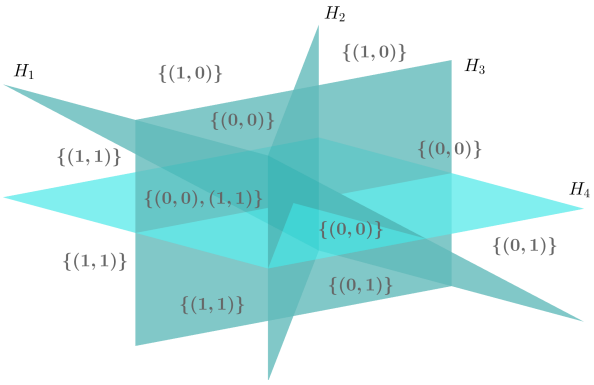
Verifying that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ requires verifying a pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ and a collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ that satisfy the conditions in Definition 2.2, an infinite-dimensional existence problem. The next result shows that the infinite-dimensional existence problem can be replaced with an equivalent finite-dimensional problem.

Theorem 2.1. *Suppose Assumptions 2.1 and 2.2 hold, let $\boldsymbol{u}_\ell \in \mathcal{U}_\ell$ be any point from the set $\mathcal{U}_\ell$, and let $\boldsymbol{U}^\dagger$ be a random vector with support on $\mathcal{U}^\dagger := \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L\}$. Then $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ if and only if there exists a $\theta \in \Theta$, a collection of finite-dimensional conditional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ that satisfy the conditions:*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} \in B \mid \boldsymbol{X} = \boldsymbol{x}\right) = \int_{\mathcal{U}^\dagger} \int_B \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{2.13}$$



(a) The figure illustrates 11 of 14 cells induced by the hyperplanes H_1, H_2, H_3 and H_4. Cells $\mathcal{U}_{12}, \mathcal{U}_{13}$ and $\mathcal{U}_{14}$ are hidden behind the figure.



(b) The figure displays the $\boldsymbol{Y}-$level sets associated with each cell in the arrangement from Figure 1(a).

Figure 1

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}\left(\boldsymbol{Y}_\gamma \in B \mid \boldsymbol{X}=\boldsymbol{x}\right) = \int_{\mathcal{U}^\dagger}\int_{B} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)\}\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{2.14}$$

$P_{\boldsymbol{X}}-$almost surely for all $B \subset \mathcal{Y}$.

The forward direction of Theorem 2.1 asserts that, if $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$, then we can construct finite-dimensional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ satisfying (2.13) and (2.14). In other words, every $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ can be rationalized by a finitely supported vector of latent variables. For the reverse direction, if there exist finite-dimensional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ that satisfy (2.13) and (2.14), then Theorem 2.1 asserts that the corresponding $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ from (2.14) belongs to the identified set $\mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. In other words, every finitely supported latent variable satisfying (2.13) and (2.14) generates a $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$.

The result suggests that for the purpose of constructing sharp bounds on counterfactual functionals it is without loss of generality to assume the latent variables have finite support given by the points $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L$ selected arbitrarily from each cell. Intuitively, the result reflects the fact that Assumptions 2.1 and 2.2 place no restrictions on the latent variable distribution *within* a cell, but instead only places constraints on the latent variable distribution *across* the cells. If there are no constraints on the latent variable distribution within a cell, we are free to distribute the mass inside a cell in any way that is convenient. In particular, the mass of the latent variable distribution inside a cell can be collapsed to a single point, and Theorem 2.1 shows it can be done without any loss of identifying information.

Each point $\boldsymbol{u}_\ell$ selected from the cell $\mathcal{U}_\ell$ is called a *witness point*, and witness points for the cells from Figure 1 are displayed in Figure 2. Since there are a finite number of cells the distribution of a latent variable with support on the witness points can be characterized by a probability vector of finite length. This is important to obtain a tractable characterization of the sharp bounds on counterfactual functionals, presented in the next subsection. There is no restriction on how a witness point is selected from a cell, and the collection of witness points $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L$ is not unique. This does not affect Theorem 2.1, which holds for any collection of witness points. Also, the latent space cells and the corresponding collection of

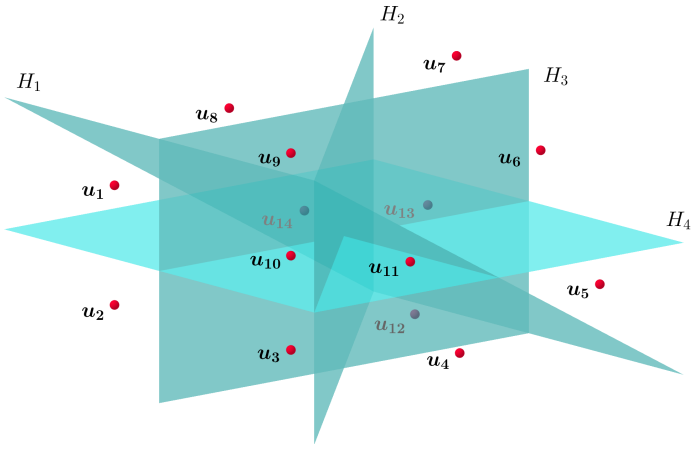


Figure 2: The figure displays the hyperplane arrangement for the game example from Figure 1 with a red witness point illustrated in each of the 14 cells. Theorem 2.1 shows that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ if and only if there exists a finite-valued random variable $\boldsymbol{U}^\dagger$ with support on $\mathcal{U}^\dagger := \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{14}\}$ satisfying conditions (2.13) and (2.14).

witness points do not depend on the counterfactual γ of interest, so long as the counterfactual is of the form specified prior to Assumption 2.2. This is by construction of our class of counterfactuals. This class of counterfactuals cannot influence the latent space cell structure in the precise sense that $\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)$ for $\boldsymbol{u} \in \mathcal{U}_\ell \iff \boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)$. In principle our method might be extended to allow for a more general class of counterfactuals that may influence the latent space cell structure, although for simplicity we do not pursue this here.

Theorem 2.1 is intuitively similar to the subdistribution extension results in Torgovitsky (2019b), but focuses on counterfactual distributions and does not require that the cells are hyper-rectangles. This is important for accommodating multidimensional and nonseparable latent variables, such as random coefficients, which do not produce latent space cells with a product structure. Gu and Russell (2021) present a similar result for a binary choice model. Unlike the results in both Torgovitsky (2019b) and Gu and Russell (2021), Theorem 2.1 accommodates nonlinear simultaneous equation models which may be incomplete and incoherent.

2.3 Optimization Formulation

Using Theorem 2.1, we now demonstrate that counterfactual functionals of the form (2.9) can be bounded by solving a sequence of linear programming problems. Fix $\theta \in \Theta$, and define the parameter vector $\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}$ with typical element labelled as $\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell)$.[17] The vector $\boldsymbol{\pi}$ is the variable over which we optimize in our result ahead, and the element $\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell)$ represents the probability:

$$P_{\boldsymbol{Y}, \boldsymbol{Y}_\gamma, \boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Y}_\gamma = \boldsymbol{y}^\star, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}).$$

[17] Note that $d_\pi = |\mathcal{X}| \cdot |\mathcal{Y}|^2 \cdot L$. Also note that $\boldsymbol{\pi}$ implicitly depends on θ, but only through the collection $\mathscr{C}$. The dependence has been suppressed in the notation.

Without loss of generality, assume that every $\boldsymbol{x} \in \mathcal{X}$ occurs with positive probability. From condition (2.13) in Theorem 2.1, we have the constraints:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}\right) = \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell), \tag{2.15}$$

for all $\boldsymbol{x} \in \mathcal{X}$. We also require the non-negativity constraints:

$$\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \geq 0, \tag{2.16}$$

for all $(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell)$ and for all $\boldsymbol{x} \in \mathcal{X}$, and the adding up conditions:

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) = 1, \qquad \sum_{y^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) = 1, \tag{2.17}$$

for all $\boldsymbol{x} \in \mathcal{X}$. We are now ready to state the main result for the section.

Theorem 2.2. *Under Assumptions 2.1 and 2.2 the identified set for the counterfactual functional $\Psi_{\boldsymbol{x}}$ is given by:*

$$\bigcup_{\theta \in \Theta} [\Psi_{\boldsymbol{x}}^{\ell b}(\theta), \Psi_{\boldsymbol{x}}^{ub}(\theta)], \tag{2.18}$$

where $\Psi_{\boldsymbol{x}}^{\ell b}(\theta)$ and $\Psi_{\boldsymbol{x}}^{ub}(\theta)$ are determined by the optimization problems:

$$\begin{aligned} \Psi_{\boldsymbol{x}}^{\ell b}(\theta) := \min_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} & \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell), \\ & \textit{s.t. } (2.15), (2.16), \textit{ and } (2.17), \end{aligned} \tag{2.19}$$

$$\begin{aligned} \Psi_{\boldsymbol{x}}^{ub}(\theta) := \max_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} & \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell), \\ & \textit{s.t. } (2.15), (2.16), \textit{ and } (2.17). \end{aligned} \tag{2.20}$$

Theorem 2.2 asserts that, given the set of points $\{\boldsymbol{u}_\ell\}_{\ell=1}^{L}$ from Theorem 2.1, the identified set for the counterfactual functional $\Psi_{\boldsymbol{x}}$ is given by the union (over all $\theta \in \Theta$) of intervals whose endpoints are determined by the value functions of the linear programs (2.19) and (2.20). While the result assumes that the distribution $P_{\boldsymbol{Y}|\boldsymbol{X}}$ is known, in Appendix E we discuss a consistency result of Gu and Russell (2021) when $P_{\boldsymbol{Y}|\boldsymbol{X}}$ is estimated which is implemented in our application. Inspection of the constraints for the linear programs (2.19) and (2.20) also shows how incoherency is addressed: since the constraints ensure that $\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) = 0$ if either $\boldsymbol{y} \notin \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)$ or $\boldsymbol{y}^\star \notin \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)$, zero probability is assigned to regions of

the latent variable space that lead to either factual or counterfactual incoherency.[18, 19] In Appendix C we discuss simplifications that arise when the model is complete.

Nevertheless, Theorem 2.2 is of limited practical use unless the researcher has an efficient method of determining the set of points $\{\boldsymbol{u}_\ell\}_{\ell=1}^L$ that satisfy the requirements of Theorem 2.1. It is also unreasonable to require that the researcher to evaluate the linear programs (2.19) and (2.20) at every $\theta \in \Theta$.[20] In the sections to come we show how to address both of these computational concerns. In particular, in the next section we present a method for enumerating a set of points $\{\boldsymbol{u}_\ell\}_{\ell=1}^L$ satisfying the requirements of Theorem 2.1 using a new cell enumeration algorithm for hyperplane arrangements. Under additional assumptions, in Section 4 we also present a method of computing the exact (i.e. not approximate) identified set using Theorem 2.2 that does not require the researcher to evaluate the linear programs (2.19) and (2.20) at every $\theta \in \Theta$.[21]

Remark 2.3. *Although we focus on counterfactual parameters, our approach can also be used to construct the identified set for structural parameters, which is given by all vectors $\theta \in \Theta$ leading to feasible linear programs in (2.19) and (2.20).*

3 Computation

3.1 The Latent Space Enumeration Algorithm

3.1.1 Algorithm Overview

For Theorem 2.2 to be practically useful the researcher must determine the set $\mathcal{U}^\dagger = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L\}$ satisfying the conditions in Theorem 2.1. The special structure of the class of models satisfying Assumption 2.1 allows us to accomplish the task using cell enumeration algorithms for hyperplane arrangements. Previous such algorithms have been used in Gu and Koenker (2020) in the context of nonparametric maximum likelihood

[18] For instance, using (2.17):

$$\sum_{\boldsymbol{y}\in\mathcal{Y}}\sum_{\ell=1}^{L}\sum_{\boldsymbol{y}^\star\in\mathcal{Y}}\pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell) - \sum_{\boldsymbol{y}^\star\in\mathcal{Y}}\sum_{\ell=1}^{L}\sum_{\boldsymbol{y}\in\mathcal{Y}}\mathbb{1}\{\boldsymbol{y}^\star\in\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_\ell,\theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell) = 0,$$

$$\iff \sum_{\boldsymbol{y}^\star\in\mathcal{Y}}\sum_{\ell=1}^{L}\sum_{\boldsymbol{y}\in\mathcal{Y}}\mathbb{1}\{\boldsymbol{y}^\star\notin\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_\ell,\theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell) = 0.$$

This and non-negativity from (2.16) ensure that $\pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell) = 0$ if $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_\ell,\theta) = \varnothing$.

[19] This approach corresponds to "IC2" in Section 4.2 in Chesher and Rosen (2020). Section 4 in Chesher and Rosen (2020) also discusses other approaches to incoherency.

[20] Instead the researcher can establish a finite grid $\Theta^\dagger \subset \Theta$ and hope to approximate the identified set by evaluating the linear programs (2.19) and (2.20) over the grid. Similar procedures have been implicitly endorsed by many applications in the literature on partial identification. Such a procedure always generates an inner approximation to the identified set, and typically is infeasible to implement when θ has more than a few dimensions.

[21] The idea is related to Proposition 2 in Torgovitsky (2019b), who demonstrates that two parameters θ and θ' that generate the same partition of the latent variable space (up to relabelling) also generate equivalent optimization problems of the form (2.19) and (2.20), so that solving these problems for both θ and θ' is redundant. In this way, it is possible to partition Θ into equivalence classes. From Torgovitsky (2019b) p. 132: *"An algorithm that could determine the partition ex ante would reap enormous computational gains, perhaps breaking the curse of dimensionality associated with a grid search. Constructing such an algorithm is not a straightforward problem... but may be a good avenue for future research."*

for random coefficient binary response models, and more recently in Gu and Russell (2021) as a method of computing bounds on counterfactual quantities in a nonseparable binary response models. Here we present a new cell enumeration algorithm. We then discuss efficient methods of profiling $\theta \in \Theta$ from the bounds in Theorem 2.2.

Given an arrangement $\mathscr{H}$, our cell enumeration procedure returns a witness point from each cell and a unique sign vector in $\{-1,1\}^{n_h}$ which indicates the position of the cell relative to the hyperplanes in the arrangement.[22] The number of cells depends on the number of hyperplanes, the dimension of the latent variable space, and whether the hyperplanes are in *general position*. The arrangement $\mathscr{H}$ is in general position if every set of k hyperplanes in $\mathscr{H}$ have intersection of dimension $d_u - k$ for $1 < k \leq d_u$ and empty intersection for $k > d_u$. An arrangement that is not in general position is *degenerate*. Intuitively, general position requires that no hyperplanes are parallel to each other and that they do not intersect in non-generic ways, maximizing the number of cells. The earliest result on the number of cells in a hyperplane arrangement was given by Buck (1943), who found that the number of cells induced by n_h hyperplanes in general position in $\mathbb{R}^{d_u}$ is:

$$\sum_{k=0}^{d_u} \binom{n_h}{k}. \tag{3.1}$$

For large n_h and fixed d_u the expression in (3.1) is $O(n_h^{d_u})$, showing that the number of cells is bounded by a polynomial in the number of hyperplanes. A hyperplane arrangement is *central* if all hyperplanes in the arrangement pass through a single point, and a central arrangement is *linear* if that single point is the origin.[23] For linear hyperplane arrangements, Cover (1965) showed that the number of cells induced by n_h hyperplanes in general position is given by:

$$2 \sum_{k=0}^{d_u-1} \binom{n_h - 1}{k}. \tag{3.2}$$

Our proposed algorithm for enumerating cells operates on linear arrangements, so we assume $\mathscr{H}$ is a linear arrangement throughout.[24]

For large n_h and fixed d_u the expression in (3.2) is $O(n_h^{d_u-1})$.

The number of cells characterized by (3.1) and (3.2) indicates that enumeration can be computationally challenging when n_h and d_u are large. The literature on computational geometry provides two main approaches: the reverse search method, introduced by Avis and Fukuda (1996), and the incremental approach,

[22] For instance, a cell represented by the sign vector $\boldsymbol{s}^\top = (-1,1,1)$ lies in the negative half-space of the first hyperplane, and the positive half-space for the second and third hyperplane. Given a witness point, a sign vector can always be computed by taking inner products between the witness points and the normal vectors of all hyperplanes in the arrangement.

[23] See Stanley (2004) p.5. Some works call an arrangement central if all hyperplanes pass through the origin (c.f. Aguiar and Mahajan (2017) p.4).

[24] This is without loss of generality since the cells induced by any hyperplane arrangement can be enumerated by use of an equivalent linear arrangement. In particular, consider a hyperplane arrangement $\{\{\boldsymbol{u} : \boldsymbol{h}_i^\top \boldsymbol{u} = b_i\} : i = 1,\ldots,n_h\}$. Each hyperplane $\{\boldsymbol{u} : \boldsymbol{h}_i^\top \boldsymbol{u} = b_i\}$ can be rewritten as $\{(\boldsymbol{u},v) : \boldsymbol{h}_i^\top \boldsymbol{u} - b_i v = 0\} \cap \{(\boldsymbol{u},v) : v = 1\}$. We can enumerate the cells of a linear arrangement with the hyperplanes $\{\{(\boldsymbol{u},v) : \boldsymbol{h}_i^\top \boldsymbol{u} - b_i v = 0\} : i = 1,\ldots,n_h\} \cup \{\{(\boldsymbol{u},v) : v = 0\}\}$. Now consider the subset of sign vectors with a last entry of "+1" for this arrangement. Dropping the last entry of each of these sign vectors, we can recover the sign vector for the original (i.e. not linear) arrangement.

which includes the algorithms of Rada and Černý (2018) and Gu and Koenker (2020). Both Rada and Černý (2018) and Gu and Koenker (2020) use linear programs to locate the new witness points as hyperplanes are added. In particular, when a new hyperplane is introduced into the arrangement, Rada and Černý (2018) first update the existing set of sign vectors. This is done by determining the sign of the existing witness points evaluated at the newly added hyperplane. The existing sign vectors are then appended with a new element—either -1 or $+1$—depending on the result. To check for newly created cells the last entry of each sign vector is flipped (i.e. multiplied by -1), creating a new set of sign vectors. A linear program can be used to check if each new sign vector corresponds to one of the newly created cells.[25] The procedure is then repeated until all hyperplanes have been added to the arrangement.

Although each linear program can be efficiently solved the total number of linear programs makes these algorithms computationally burdensome. Our proposed algorithm is also incremental, but exploits the recursive structure of hyperplane arrangements. This allows us to reduce the arrangement problem in higher dimensions to a series of simpler arrangements in $\mathbb{R}^2$ where we can perform enumeration without the need to solve any linear programs. The algorithm can also easily handle degeneracy, a common occurrence in the models we consider.

The key challenge with incremental algorithms is to find the witness points associated with each newly created cell. To illustrate how our algorithm locates these cells, consider the arrangement $\mathscr{H} = \{H_1, \ldots, H_{n_h}\}$, which consists of the hyperplane H_{n_h} added to the subarrangement $\mathscr{H}' = \{H_1, \ldots, H_{n_h-1}\}$. Furthermore, let $\mathscr{H}'' = \{H_1 \cap H_{n_h}, \ldots, H_{n_h-1} \cap H_{n_h}\}$ be the lower-dimensional arrangement on H_{n_h} consisting of the intersection of H_{n_h} with the $n_h - 1$ hyperplanes in $\mathscr{H}'$. Note that when H_{n_h} is added to $\mathscr{H}'$ it will split some of the existing cells to create new cells. Now every cell $\mathcal{C}$ crossed by H_{n_h} is uniquely associated with a lower-dimensional cell $\mathcal{C} \cap H_{n_h}$ induced by the arrangement $\mathscr{H}''$. The witness points from the lower-dimensional arrangement $\mathscr{H}''$ lie within the crossed cells. Perturbing the witness points from $\mathscr{H}''$ in a suitable way then allows us to generate witness points in the new cells created by the addition of H_{n_h}. This effectively reduces the problem of finding witness points in d_u dimensions using the arrangement $\mathscr{H}$ into the problem of finding witness points in $d_u - 1$ dimensions using the arrangement $\mathscr{H}''$. Now $\mathscr{H}''$ is itself an arrangement of $n_h - 1$ hyperplanes in $d_u - 1$ dimensions, and this procedure can be repeated on the arrangement $\mathscr{H}''$. Our algorithm then uses recursion on this relation between $\mathscr{H}$, $\mathscr{H}'$ and $\mathscr{H}''$ to decompose every higher-dimensional arrangement into a sequence of two dimensional arrangements, where witness points can be found without using linear programs. The combination of recursion with a cheap two-dimensional base case makes our algorithm substantially faster than previous state-of-the-art cell enumeration algorithms.

Our algorithm also connects to a well-known result in the literature on hyperplane arrangements: if

[25] In particular, consider a $n_h \times d_u$ matrix $\boldsymbol{H}$ with rows consisting of the normal vectors of the hyperplanes in our arrangement. Now for a fixed sign vector $\boldsymbol{s} \in \{-1, +1\}^{n_h}$, consider the linear program:

$$\varepsilon^* = \max_{\varepsilon, \boldsymbol{u}} \varepsilon \quad \text{s.t.} \quad \boldsymbol{s} \odot \boldsymbol{H}\boldsymbol{u} \geq \varepsilon,$$

where $\odot$ denotes the elementwise product. Then there exists a cell with nonempty interior defined by a sign vector $\boldsymbol{s} \in \{-1, +1\}^{n_h}$ if and only if $\varepsilon^* > 0$.

$r(\mathscr{H}_0)$ denotes the number of cells induced by an arrangement $\mathscr{H}_0$, then:[26]

$$r(\mathscr{H}) = r(\mathscr{H}') + r(\mathscr{H}'').$$

This result can be used to prove that our algorithm finds witness points from all cells in any linear arrangement.

Example 2 (Cont'd)**.** *Consider again the game example with two players, and consider the system of equations:*

$$Y_{i1} = \mathbb{1}\{Y_{i2}U_{i3} - U_{i1} \geq 0\},$$
$$Y_{i2} = \mathbb{1}\{Y_{i1}U_{i3} - U_{i2} \geq 0\}.$$

Recall the hyperplanes are given by:

$$H_1 = \{\boldsymbol{u} \in \mathbb{R}^3 : u_3 - u_1 = 0\}, \qquad H_3 = \{\boldsymbol{u} \in \mathbb{R}^3 : u_3 - u_2 = 0\},$$
$$H_2 = \{\boldsymbol{u} \in \mathbb{R}^3 : -u_1 = 0\}, \qquad H_4 = \{\boldsymbol{u} \in \mathbb{R}^3 : -u_2 = 0\}.$$

An illustration of our recursive cell enumeration procedure is provided in Figure 3. The hyperplanes H_1, H_2 and H_3 form the subarrangement $\mathscr{H}'$, and are colored ***black****. The final hyperplane H_4 is colored* ***blue****. The three* ***black*** *hyperplanes from Figure 3(a) representing H_1, H_2 and H_3 partition the space $\mathbb{R}^{d_u}$ into $r(\mathscr{H}') = 8$ cells. Out of the 8 cells in the subarrangement $\mathscr{H}'$, only 6 are crossed by H_4; in Figure 3(a) we highlight one of the cells not crossed by H_4 in red. Our main innovation is to find the crossed cells using the lower-dimensional arrangement $\mathscr{H}''$ illustrated in Figure 3(b).*

In Figure 3(b) the intersections of the three ***black*** *hyperplanes H_1, H_2 and H_3 with the* ***blue*** *hyperplane H_4 form three lines which partition the subspace H_4 into $r(\mathscr{H}'') = 6$ cells. These lines are:*

$$H_1 \cap H_4 = \{\boldsymbol{u} \in H_4 : -u_1 - u_2 = 0\},$$
$$H_2 \cap H_4 = \{\boldsymbol{u} \in H_4 : -u_1 = 0\},$$
$$H_3 \cap H_4 = \{\boldsymbol{u} \in H_4 : -u_2 = 0\},$$

and they represent the lower-dimensional arrangement $\mathscr{H}''$ on the hyperplane H_4.[27] *The witness points for the cells in the lower-dimensional arrangement $\mathscr{H}''$ are labelled A to F. These witness points can be found without solving any linear programs, for instance, by taking convex combinations of points lying on two adjacent hyperplanes in the lower-dimensional arrangement. A version of the procedure is implemented in Algorithm 1. By construction, these witness points are in the interior of the cells that are crossed when*

[26] The earliest reference of the relationship appears to be in Cover (1965). See also Lemma 4A1 in Zaslavsky (1975), page 4 in Orlik and Terao (1992), and Lemma 2.1 in Stanley (2004).

[27] The construction of the arrangement $\mathscr{H}''$ requires a change of basis since it needs to be expressed as an arrangement on the hyperplane H_4. For instance, let $\boldsymbol{h}_4 = (0, 0, 1)^\top$ be the normal vector of hyperplane H_4, and let $N(\boldsymbol{h}_4)$ be an orthonormal basis for the null space of $\boldsymbol{h}_4$. Then $N(\boldsymbol{h}_4)$ is an orthonormal basis that spans all elements on H_4. Then the normal vectors for $H_1 \cap H_4$, $H_2 \cap H_4$ and $H_3 \cap H_4$ are given by $(N(\boldsymbol{h}_4)^\top N(\boldsymbol{h}_4))^{-1} N(\boldsymbol{h}_4)^\top \boldsymbol{h}_1$, $(N(\boldsymbol{h}_4)^\top N(\boldsymbol{h}_4))^{-1} N(\boldsymbol{h}_4)^\top \boldsymbol{h}_2$ and $(N(\boldsymbol{h}_4)^\top N(\boldsymbol{h}_4))^{-1} N(\boldsymbol{h}_4)^\top \boldsymbol{h}_3$ where $\boldsymbol{h}_1$, $\boldsymbol{h}_2$ and $\boldsymbol{h}_3$ are normals of H_1, H_2 and H_3.

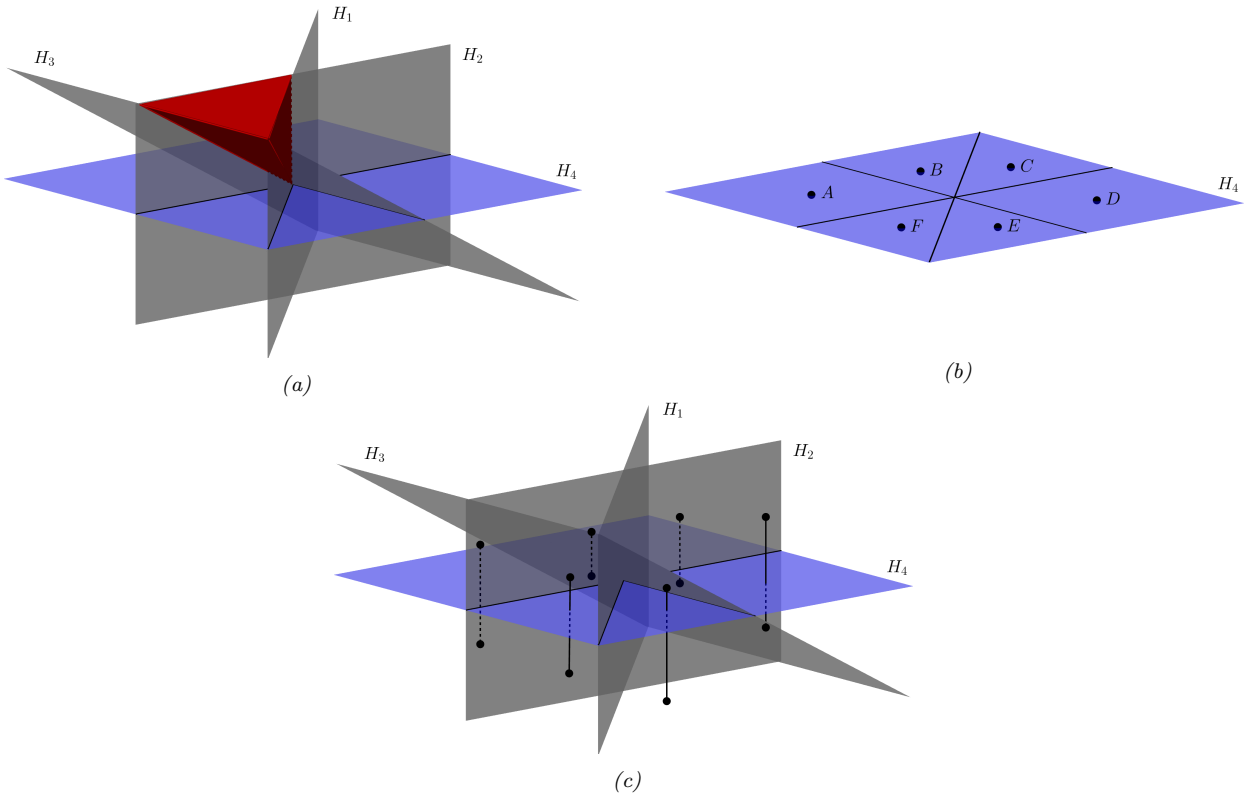


Figure 3: Figure (a) illustrates an arrangement $\mathscr{H}$ consisting of $n_h = 4$ hyperplanes in $d_u = 3$ dimensional space. The sub-arrangement $\mathscr{H}'$ consists of the three black hyperplanes, and the newly added hyperplane H_4 is in blue. The lower-dimensional arrangement $\mathscr{H}''$ is represented by the three black lines the intersection of H_1, H_2 and H_3 with H_4. Figure (a) also highlights a cell which is not crossed by H_4. The subarrangement $\mathscr{H}'$ has 8 cells, and only 6 are crossed by H_4. Figure (b) shows the six witness points found in the lower-dimensional arrangement $\mathscr{H}''$. Figure (c) illustrates our perturbation procedure to move witness points from the lower-dimensional arrangement into the newly crossed cells in the higher-dimensional arrangement.

H_4 is added, allowing us to efficiently find the crossed cells without solving any linear programs.

To locate witness points in the newly created cells, we perturb the witness points A to F slightly away from H_4, yielding 12 *new witness points in* $\mathbb{R}^3$. *This is illustrated in Figure 3(c). Details on the magnitude and direction of the perturbations are presented in Algorithm 2. After perturbation, some cells may have two witness points. We resolve this in our algorithm by caching the witness points from all previous iterations, determining which cells have been crossed by the newly added hyperplane at each iteration, discarding any cached witness points that correspond to the crossed cells, and adding the new perturbed witness points to the cache. To determine which cells have been crossed by the newly added hyperplane, we must determine the sign vector associated with each previous and newly added witness points and then discard a witness point from any pair of duplicate sign vectors. Determining the duplicate sign vectors at each iteration is called matching, and influences our time complexity result.*

In the example, it is easy to enumerate the cells induced by the lower-dimensional arrangement $\mathscr{H}''$, which forms a two-dimensional subspace. In the general case, newly added hyperplanes form $d_u - 1$ dimensional

subspaces of the d_u-dimensional ambient space, so that enumeration on $\mathscr{H}''$ is more complicated. In these cases our algorithm is called recursively, with the simple two-dimensional arrangement as the base case.

Handling degeneracy is also straightforward in our approach. When hyperplanes are not in general position in a linear arrangement the lower-dimensional arrangement is still determined by the intersections of all hyperplanes in the existing arrangement with the newly added hyperplane, although these intersections may no longer be distinct. Since all linear arrangements in dimension d_u can be recursively reduced to linear arrangements in $\mathbb{R}^2$ for the purpose of enumeration, and since degenerate linear arrangements in $\mathbb{R}^2$ consist of overlapping lines, dealing with degenerate arrangements amounts to removing duplicate hyperplanes in $\mathbb{R}^2$.

3.1.2 Algorithm Details

While the example in Figure 3 provides the intuition for our approach, Algorithms 1 and 2 provide the details. Algorithm 1 enumerates linear arrangements of dimension $d_u = 2$, which consist of lines passing through the origin. Algorithm 1 operates by ordering the lines according to their angle, and then takes the average of the points on two adjacent lines to identify a witness point for each region. Specifically, steps 4(a) and 4(b) find the points of intersection of each hyperplane with the unit circle, and step 4(c) then sorts these points using their angle in polar coordinates.[28] Steps 4(d) and 4(e) then construct the witness points by taking the midpoint of the sorted points p_i on the unit circle. Sorting the points in step 4(c) of Algorithm 1 with n_h hyperplanes has time complexity $O(n_h \log(n_h))$ in the worst case, and all other operations scale linearly with n_h. Because of this, Algorithm 1 has a time complexity $O(n_h \log(n_h))$ in the worst case.

Algorithm 2 operates by successively adding hyperplanes to a growing arrangement. The normal vectors of all previous hyperplanes are first projected onto the newly added hyperplane (step 3(c)). After the projection, we are left with a linear arrangement in dimension $d_u - 1$. At this point, Algorithm 2 is called again if $d_u - 1 > 2$, and Algorithm 1 is called if $d_u - 1 = 2$ (step 3(d)). This is the recursive component of Algorithm 2. After Algorithm 1 is called from within Algorithm 2, we perturb the witness points away from the newly added hyperplane. The size of the perturbation determined in step 3(f)(i) guarantees that, while moving away from the newly added hyperplane to find a witness point in the new cell, we do not cross any other hyperplanes into an adjacent cell.[29] After the perturbation, one of the perturbed witness points shares the same sign vector as one of the witness points from the previous iteration. After identifying which witness points share the same sign vector in the matching step (3(f)(iv)), we drop one of the matched witness points. We are then left with an updated set of witness points and sign vectors that can be fed into the subsequent iteration.

[28] In the algorithm the angle is calculated with respect to the first row of $\boldsymbol{H}$.

[29] In particular, consider step 3(f)(i) of Algorithm 2. Let $\boldsymbol{h}_i$ denote the normal vector for hyperplane H_i. Each unique hyperplane in $\{H_1, ..., H_{i-1}\}$ intersects the line expressed parametrically as $t_k \mapsto \boldsymbol{w}_k^* + t_k \boldsymbol{h}_i$ exactly once, unless it is perpendicular to H_i (in which case it imposes no constraint on how large t_k can be without crossing it). At the point t_{kj} that the line intersects the hyperplane indexed by j, we have $\boldsymbol{h}_j^\top(\boldsymbol{w}_k^* + t_{kj}\boldsymbol{h}_i) = 0$, or, equivalently, $t_{kj} = \boldsymbol{h}_j^\top \boldsymbol{w}_k^* / \boldsymbol{h}_j^\top \boldsymbol{h}_i$. Any points on the line strictly between $\boldsymbol{w}_k^* - m_k \boldsymbol{h}_i$ and $\boldsymbol{w}_k^* + m_k \boldsymbol{h}_i$ have crossed no hyperplanes and have sign vectors equal to that of $\boldsymbol{w}_k^*$ for hyperplanes up to the $i-1$th.

Algorithm 1: ($d_u = 2$)

input : $n \times 2$ matrix $\boldsymbol{H}$, where $\boldsymbol{Hu} = 0$ represents a linear arrangement.
output: Matrix of witness points $\boldsymbol{W}$ and the associated sign vectors $\boldsymbol{S}$.

1. Drop duplicated (up to a constant) rows of $\boldsymbol{H}$ and call the resulting matrix $\boldsymbol{H}^*$.
2. Let n_u be the number of rows in $\boldsymbol{H}^*$. If $n_u = 1$, proceed to step 3. Otherwise, proceed to step 4.
3. **If** $n_u = 1$:
   (a) Pick a random point $\boldsymbol{w}_1$ in $\mathbb{R}^{d_u}$.
   (b) If $\langle \boldsymbol{a}_1^*, \boldsymbol{w}_1 \rangle < 0$, set $\boldsymbol{w}_1 \leftarrow -\boldsymbol{w}_1$.
   (c) Set $\boldsymbol{W}^+ = \boldsymbol{w}_1$. Proceed to step 5.
4. **If** $n_u > 1$:
   (a) Set $\boldsymbol{p}_i = \frac{1}{||\boldsymbol{h}_i^*||} \boldsymbol{h}_i^* \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ for $i = 1, 2, \ldots, n_u$ of $\boldsymbol{H}^*$, where $\boldsymbol{h}_i^*$ is the ith row of $\boldsymbol{H}^*$.
   (b) For each i, if $\langle \boldsymbol{h}_1^*, \boldsymbol{p}_i \rangle$ is negative, set $\boldsymbol{p}_i = -\boldsymbol{p}_i$.
   (c) Order the points $\boldsymbol{p}_i$ by $\langle \boldsymbol{h}_1^*, \boldsymbol{p}_i \rangle$ (increasing), and let the ordered points be $\boldsymbol{q}_j$, with $j = 1, 2, \ldots, n_u$.
   (d) For $j \leq n_u - 1$, let $\boldsymbol{w}_j = \frac{1}{2}(\boldsymbol{q}_j + \boldsymbol{q}_{j+1})$.
   (e) Let $\boldsymbol{w}_{n_u} = \frac{1}{2}(\boldsymbol{q}_{n_u} - \boldsymbol{q}_1)$.
   (f) Let $\boldsymbol{W}^+$ be the $2 \times n_u$ matrix with points $\boldsymbol{w}_j$ as columns.
5. Let $\boldsymbol{W}$ be the matrix composed of the columns of $\boldsymbol{W}^+$ and the columns of $-\boldsymbol{W}^+$. Let $\boldsymbol{S}$ be the matrix of signs of $\boldsymbol{HW}$. Return $\boldsymbol{W}$ and $\boldsymbol{S}$.

For any $d_u > 2$ the total number of calls to Algorithm 1 in the general position case is $O(n_h^{d_u - 2})$.[30] Since Algorithm 1 has time complexity $O(n_h \log(n_h))$ the time complexity for recursively calling Algorithm 1 is $O(n_h^{d_u - 1} \log(n_h))$ in the worst case, or $O(n_c \log(n_h))$ where n_c is the upper bound of the number of cells in a linear arrangement in dimension d_u, given in (3.2). This means the time complexity spent of Algorithm 2 for recursive enumeration is nearly optimal, since it is proportional to the total number of cells in the arrangement up to a logarithm factor.

The dominant step in Algorithm 2 for complexity is matching (step 3(f)(iv)) and the time complexity for the matching step is $O(n_h n_c mat(n_h))$ where $\mathrm{mat}(n_h)$ is the time cost for determining whether two vectors of length n_h are identical. Our Algorithm 2 for cell enumeration for any $d_u > 2$ is $O(n_h n_c \mathrm{mat}(n_h))$. In comparison, the algorithm of Rada and Černý (2018) has complexity $O(n_h n_c \mathrm{lp}(n_h, d_u))$ and the algorithm of Gu and Koenker (2020) has complexity $O(n_c \mathrm{lp}(n_h, d_u) + n_h n_c \mathrm{mat}(n_h))$, where $\mathrm{lp}(n_h, d_u)$ denotes the cost

[30] Let $C_2(n, d)$ be the number of calls to Algorithm 1 for an arrangement of n hyperplanes in d dimensions. Then $C_2(n, 3) = \binom{n-1}{1}$ (for each hyperplane, except the first, there is one call to Algorithm 1), and when $d > 3$ we have $C_2(n, d) = \sum_{i=1}^{n-1} C_2(i, d-1)$ calls to Algorithm 1. Combining, this yields:

$$C_2(n_h, d_u) = \sum_{i_1=1}^{n_h - 1} \sum_{i_2=1}^{i_1 - 1} \cdots \sum_{i_{d_u-4}=1}^{i_{d_u-5}-1} \sum_{i_{d_u-3}=1}^{i_{d_u-4}-1} C_2(i_{d_u-3}, 3) = \sum_{i_1=1}^{n_h - 1} \sum_{i_2=1}^{i_1 - 1} \cdots \sum_{i_{d_u-4}=1}^{i_{d_u-5}-1} \sum_{i_{d_u-3}=1}^{i_{d_u-4}-1} \binom{i_{d_u-3}}{1} = \binom{n_h - 1}{d_u - 2} = O(n_h^{d_u - 2}).$$

Algorithm 2: ($d_u > 2$)

input : An $n_h \times d_u$ matrix $\boldsymbol{H}$, where $\boldsymbol{H}\boldsymbol{u} = 0$ represents a linear arrangement.
output: A matrix of witness points $\boldsymbol{W}$ and the matrix of sign vectors $\boldsymbol{S}$.

1. Pick a random point $\boldsymbol{w}_1$ in $\mathbb{R}^{d_u}$. If $\langle \boldsymbol{h}_1, \boldsymbol{w}_1 \rangle < 0$, set $\boldsymbol{w}_1 \leftarrow -\boldsymbol{w}_1$.
2. Store $\boldsymbol{w}_1$ as a column in the matrix $\boldsymbol{W}^+$, and initialize $\boldsymbol{S}^+$ as an $n_h \times 1$ zero matrix, then set $S_{11} = 1$.
3. For $i = 2, 3, \ldots, n_h$:
   (a) Let $\boldsymbol{s}_i = \text{sgn}(\boldsymbol{h}_i^\top \boldsymbol{W}^+)$, and store $\boldsymbol{s}_i$ as the i^{th} row of $\boldsymbol{S}^+$.
   (b) If $\boldsymbol{h}_i$ is a duplicated row (up to a constant) set $i \leftarrow i + 1$ and go to 3(a).
   (c) Let $\boldsymbol{H}_P = \boldsymbol{H}_{1:(i-1)} N(\boldsymbol{h}_i)$, where $N(\,\cdot\,)$ denotes a $d_u \times (d_u - 1)$ orthonormal matrix with columns given by a basis of the orthogonal complement of $\boldsymbol{h}_i$.
   (d) If $d_u - 1 > 2$, call Algorithm 2 (this algorithm) recursively on $\boldsymbol{H}_P$. Otherwise, call Algorithm 1. Let $\boldsymbol{W}_P$ denote the returned witness points.
   (e) Set $\boldsymbol{W}^* = N(\boldsymbol{h}_i)\boldsymbol{W}_P$, where $N(\,\cdot\,)$ denotes a $d_u \times (d_u - 1)$ matrix with columns given by a basis of the orthogonal complement of $\boldsymbol{h}_i$.
   (f) Let J_i be the number of columns in $\boldsymbol{W}^*$. For $k = 1, 2, \ldots, J_i$:
      i. Let m_k be some number in $(0, \min_{j=1\ldots i-1} \left| \frac{\langle \boldsymbol{h}_j, \boldsymbol{w}_k^* \rangle}{\langle \boldsymbol{h}_j, \boldsymbol{h}_i \rangle} \right|)$, where $\boldsymbol{w}_k^*$ is the k^{th} column of $\boldsymbol{W}^*$.
      ii. Add and substract $m_k \boldsymbol{h}_i$ to the kth column of $\boldsymbol{W}^*$ to create two new vectors $\tilde{\boldsymbol{w}}_1$ and $\tilde{\boldsymbol{w}}_2$.
      iii. Find the sign vectors $\tilde{\boldsymbol{s}}_1 := \boldsymbol{H}_{1:(i-1)}\tilde{\boldsymbol{w}}_1$ and $\tilde{\boldsymbol{s}}_2 := \boldsymbol{H}_{1:(i-1)}\tilde{\boldsymbol{w}}_2$.
      iv. Determine the column in $\boldsymbol{W}^+$ whose sign vector matches either $\tilde{\boldsymbol{s}}_1$ or $\tilde{\boldsymbol{s}}_2$, and discard it from $\boldsymbol{W}^+$. Also discard its sign vector from $\boldsymbol{S}^+$.
      v. Redefine $\boldsymbol{W}^+ \leftarrow \text{concatenate}(\boldsymbol{W}^+, \tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2)$ (column-wise) and $\boldsymbol{S}^+ \leftarrow \text{concatenate}(\boldsymbol{S}^+, \tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2)$ (column-wise).
4. Let $\boldsymbol{W}$ be the matrix composed of the columns of $\boldsymbol{W}^+$ and $-\boldsymbol{W}^+$, and let $\boldsymbol{S}$ be the matrix composed of the columns of $\boldsymbol{S}^+$ and $-\boldsymbol{S}^+$. Return $\boldsymbol{W}$ and $\boldsymbol{S}$.

of solving a linear program with n_h inequalities in d_u dimensions. For reference, the implementation of matching used in R has a time complexity of $O(n_h)$, where solving a linear program in R using Rmosek has a worst case time complexity of $O(n_h^{3.5}L)$, where L is the number of bits needed to store the constraints and objective function.[31]

To practically demonstrate the improvements made by our proposed algorithm, we compare the time performance to the algorithms in Rada and Černý (2018) and Gu and Koenker (2020) on a set of simulated hyperplanes in two different designs.[32] In both designs we generate a $n_h \times d_u$ matrix $\boldsymbol{H}$ with rows representing the normal vectors for each hyperplane in our arrangement. In the first design the entries of $\boldsymbol{H}$ are randomly generated from a standard normal distribution: the design corresponds to the case of a general position linear arrangement, and we expect the number of cells to be equal to the upper bound in (3.2). In the second design, we generate the rows of $\boldsymbol{H}$ to correspond to a degenerate arrangement. In particular, looping over all pairs $(j, k) \in \{1, \ldots, d_u\}$, a typical row in $\boldsymbol{H}$ is constructed to have 1 in the j^{th} position, -1 in the k^{th}

[31] See Andersen (2009).
[32] For both designs, the number of cells is known a priori, so that we are able to check whether the algorithms find all cells.

Table 2: Computation time in seconds of the proposed algorithm (denoted GRS) compared to the algorithms of Gu and Koenker (2020) (GK) and Rada and Černý (2018) (RC), which were implemented using the `RCBR` package in R. The reported times are median times across 3 runs. Cases that were infeasible with available computing resources have time ∞. The GK and RC algorithms are identical when $d_u > 2$.

Design 1 (general position)

	$d_u = 2$			$d_u = 3$		$d_u = 4$	
n_h	GRS	GK	RC	GRS	GK/RC	GRS	GK/RC
25	< 0.01	0.16	1.15	< 0.01	9.42	0.07	49.9
50	< 0.01	0.52	2.21	0.02	72.4	0.36	432
100	< 0.01	1.37	9.96	0.06	723	2.75	8198
200	< 0.01	6.66	51.7	0.35	7336	30.4	∞
400	0.01	64.2	325	2.29	∞	477	∞
800	0.04	851	2791	19.14	∞	8221	∞

Design 2 (degenerate)

$d_u(n_h)$	GRS	GK/RC
4 (6)	< 0.01	0.10
5 (10)	0.01	0.57
6 (15)	0.06	4.39
7 (21)	0.44	38
8 (28)	3.55	367
9 (36)	33	4127
10 (45)	334	∞

position and 0 in all other positions, for a total of $n_h = \binom{d_u}{2}$ rows. In this case there are $d_u!$ cells, which is much smaller than the upper bound in (3.2).

The simulation results are displayed in Table 2. In both designs, all three algorithms find the correct number of cells, but differ dramatically in their computation time. The improvement in Gu and Koenker (2020) in comparison to RC is noticeable for dimension two, demonstrating that the cost of solving many linear programs can be significant. For $d_u > 2$ the number of linear programs to solve can be overwhelming. In Design 1, for some of the large (n_h, d_u) combination, it becomes prohibitive to run the Gu and Koenker (2020) and Rada and Černý (2018) algorithms, while our proposed algorithm—which avoids linear programs—is feasible and fast for all (n_h, d_u) combinations.

3.2 The Profiling Algorithm

Recall from the union in (2.18) from Theorem 2.2 that we must solve a pair of linear programming problems at every possible vector of structural parameters. The procedure is infeasible in practice, and the researcher must instead solve these programs at each point on a grid over the parameter space. In most cases the grid-based approach to profiling is computationally costly and delivers only an (inner) approximation to the sample analog of the identified set.

As initially shown in Gu and Russell (2021) for a binary response model, when the index functions are linear in the structural parameters it is possible to compute the exact identified set by first determining a finite set of *representative points* in the parameter space, then solving the linear programs only at these representative points. Here we show that the intuition in Gu and Russell (2021) can be extended to the more general setting considered in this paper. However, the algorithm in Gu and Russell (2021) becomes prohibitive when n_h or d_u is large. We thus propose a new algorithm that is substantially faster than the one proposed in Gu and Russell (2021).

To begin, we formally introduce our assumption of linearity in parameters, which is required for our profiling procedure.

Assumption 3.1 (Linearity in Parameters)**.** *For all $m_1 \in \mathcal{M}_1$, $m_2 \in \mathcal{M}_2$, $k = 1, \ldots, d_y$, $A \in \mathcal{A}_k^{(m_1,m_2)}$, and all $j \in A$ the index functions $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \cdot) : \mathbb{R}^{d_u} \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ are jointly linear in $(\boldsymbol{u}, \theta)$ for each fixed $(\boldsymbol{y}, \boldsymbol{x})$.*

Notice that under Assumption 3.1 the sets:

$$\left\{(\boldsymbol{u}, \theta) \in \mathbb{R}^{d_u + d_\theta} : g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\right\}, \tag{3.3}$$

are hyperplanes in $\mathbb{R}^d$, where $d := d_u + d_\theta$. The first main idea behind our profiling procedure is to use the hyperplane arrangement formed from the hyperplanes in (3.3) to divide $\mathbb{R}^{d_\theta}$ into equivalence classes.

Definition 3.1 (Equivalence Relation)**.** *Let $\boldsymbol{A} = [\boldsymbol{A}_1 \ \boldsymbol{A}_2]$ be a $n \times (d_u + d_\theta)$ matrix with rows given by the normal vectors of the hyperplanes in a linear arrangement $\mathscr{A}$, where $\boldsymbol{A}_1$ is $n \times d_u$ and $\boldsymbol{A}_2$ is $n \times d_\theta$. For any $\boldsymbol{s} \in \{-1, +1\}^n$, let:*

$$\mathcal{U}(\theta, \boldsymbol{s}) = \{\boldsymbol{u} \in \mathcal{U} : sgn(\boldsymbol{A}_1 \boldsymbol{u} + \boldsymbol{A}_2 \theta) = \boldsymbol{s}\},$$

and let $\mathcal{S}(\theta) = \{\boldsymbol{s} \in \{-1, +1\}^n : \mathcal{U}(\theta, \boldsymbol{s}) \neq \varnothing\}$. We say θ is equivalent to θ', and write $\theta \sim \theta'$, if $\mathcal{S}(\theta) = \mathcal{S}(\theta')$.

The equivalence relation from Definition 3.1 can be used to divide Θ into equivalence classes $\Theta/\sim$. Any $\theta, \theta' \in \Theta$ belonging to the same equivalence class are associated with the same collection of latent space cells, which are uniquely characterized by their sign vectors. The following proposition also asserts that the $\boldsymbol{Y}-$level sets for θ are the same as the $\boldsymbol{Y}-$level sets for θ'.

Proposition 3.1. *Suppose Assumptions 2.1, 2.2, and 3.1 hold. For any $\theta, \theta' \in \Theta$, if $\mathcal{S}(\theta) = \mathcal{S}(\theta')$ then:*

$$\{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta))\}_{\boldsymbol{u} \in \mathcal{U}} = \{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta'), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta'))\}_{\boldsymbol{u} \in \mathcal{U}},$$

for all $\boldsymbol{x} \in \mathcal{X}$. Furthermore, if $\mathcal{S}(\theta) \subset \mathcal{S}(\theta')$ then:

$$\{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta))\}_{\boldsymbol{u} \in \mathcal{U}} \subseteq \{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta'), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta'))\}_{\boldsymbol{u} \in \mathcal{U}},$$

for all $\boldsymbol{x} \in \mathcal{X}$.

To appreciate the result, note that the constraints in the linear programs of Theorem 2.1 are only affected by θ through the collection:

$$\{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta))\}_{\boldsymbol{u} \in \mathcal{U}} = \{(\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta))\}_{\ell=1}^L. \tag{3.4}$$

It follows that any two linear programs associated with the same collection from (3.4) will have the same feasible region and objective function. By Proposition 3.1, if $\theta \sim \theta'$ then both θ and θ' yield the same collections in (3.4), and so yield equivalent linear programming problems in Theorem 2.2. It follows that the linear programs of Theorem 2.2 need only be solved at one point from each equivalence class, called a *representative point*. The second part of Proposition 3.1 shows that if $\mathcal{S}(\theta) \subset \mathcal{S}(\theta')$, then the possible

$\boldsymbol{Y}-$level sets at θ are contained in the possible $\boldsymbol{Y}-$level sets at θ'. In this case, the value functions from Theorem 2.2 must be such that $[\Psi_{\boldsymbol{x}}^{\ell b}(\theta), \Psi_{\boldsymbol{x}}^{ub}(\theta)] \subseteq [\Psi_{\boldsymbol{x}}^{\ell b}(\theta'), \Psi_{\boldsymbol{x}}^{ub}(\theta')]$, so that the linear programs in Theorem 2.2 do not need to be solved at θ if they are solved at θ' when computing the bounds using (2.18). We return to this point at the end of this section.

Now note that the set $\mathcal{S}(\theta)$ from Definition 3.1 can be rewritten as:

$$\mathcal{S}(\theta) = \{\boldsymbol{s} \in \{-1, +1\}^n : \theta \in \Theta(\boldsymbol{s})\}\,, \text{ where } \Theta(\boldsymbol{s}) := \{\theta \in \Theta : \exists \boldsymbol{u} \text{ s.t. } \operatorname{sgn}(\boldsymbol{A}_1\boldsymbol{u} + \boldsymbol{A}_2\theta) = \boldsymbol{s}\}\,. \tag{3.5}$$

Here $\Theta(\boldsymbol{s})$ is exactly the projection of the cell in $\mathcal{U} \times \Theta$ with sign vector $\boldsymbol{s}$ onto Θ. This demonstrates that $\theta \sim \theta'$ if θ and θ' belong in the intersection of the same set of projected cells from $\mathcal{U} \times \Theta$. The connection between the equivalence classes and the boundaries of $\text{Proj}_{\Theta}(\mathcal{C})$, the projection of a cell $\mathcal{C}$ on $\mathbb{R}^{d_\theta}$, is described formally in the following result.

Proposition 3.2. *Consider the linear arrangement $\mathscr{A}$. Then $\theta, \theta' \in \mathbb{R}^{d_\theta}$ satisfy $\mathcal{S}(\theta) \neq \mathcal{S}(\theta')$ if and only if there exists a cell $\mathcal{C}$ induced by $\mathscr{A}$ such that either:*

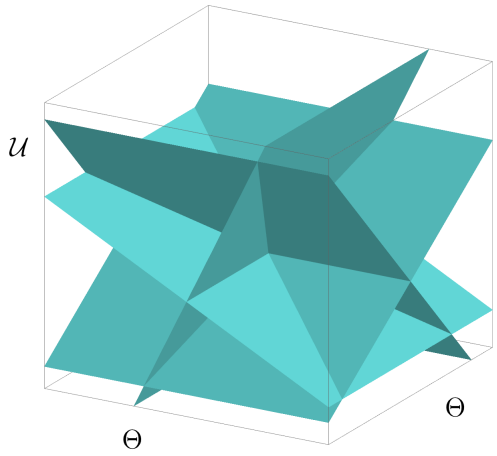
(i) $\theta \in Proj_{\Theta}(\mathcal{C})$ and $\theta' \notin Proj_{\Theta}(\mathcal{C})$, or

(ii) $\theta \notin Proj_{\Theta}(\mathcal{C})$ and $\theta' \in Proj_{\Theta}(\mathcal{C})$.

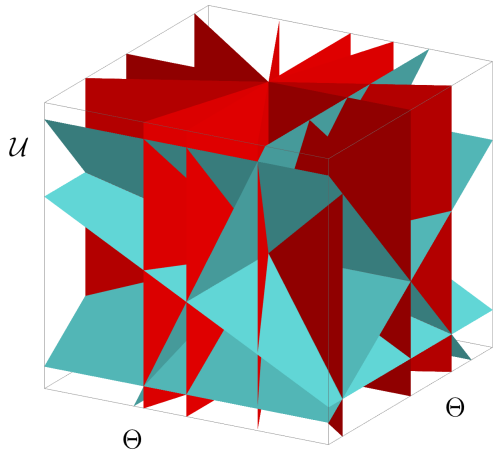
Moving over the parameter space Θ, Proposition 3.2 states that it is impossible to "cross" into a new equivalence class without also either "entering" or "leaving" the projection of some cell, and vice versa. The result implies the boundaries of the projected cells are also the boundaries of the equivalence classes. This observation motivated the profiling algorithm of Gu and Russell (2021). The first step in the profiling algorithm of Gu and Russell (2021) was to consider a linear hyperplane arrangement in $\mathbb{R}^d$ defined by the hyperplanes of the form (3.3). An example of such an arrangement is illustrated in Figure 4(a). After enumerating cells in this higher-dimensional arrangement, the cells are projected onto $\mathbb{R}^{d_\theta}$. Representative points were then taken as witness points from an application of the cell enumeration algorithm to the boundaries of the projected cells in $\mathbb{R}^{d_\theta}$, illustrated in Figure 4(d). The procedure of Gu and Russell (2021) avoided the need to blindly grid over the parameter space to construct the identified set, and can be used to compute exact (i.e. not approximate) bounds on counterfactual outcomes. Despite vast improvements over a blind gridding approach, the profiling algorithm of Gu and Russell (2021) struggles as the dimension and number of hyperplanes increase. Our new algorithm returns the same set of projected cell boundaries as in Gu and Russell (2021), but replaces the expensive higher-dimensional enumeration and projection step with a sequence of Gaussian elimination (row reduction) steps.

Our new procedure is presented in Algorithm 3. To describe the new approach, we first define a few key objects. The next two definitions are standard from the convex analysis literature, and are presented here for completeness.

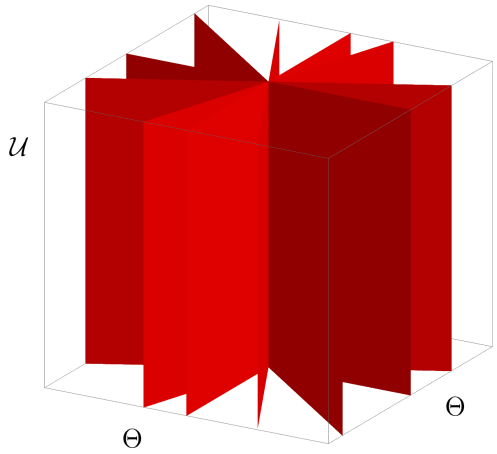
Definition 3.2 (Faces)**.** *Let $\mathcal{V} := \left\{\boldsymbol{v} \in \mathbb{R}^k : \boldsymbol{B}\boldsymbol{v} \leq 0\right\}$ be a nonempty polyhedral set, where $\boldsymbol{B}$ is an $n \times k$*



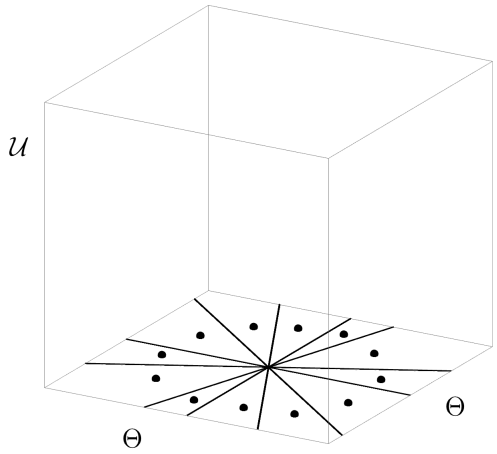
(a) A hyperplane arrangement in $\mathbb{R}^d$, with $d = d_u + d_\theta$.



(b) The arrangement from 4(a), and all profiling hyperplanes.



(c) All profiling hyperplanes for the arrangement from 4(a).



(d) The boundaries of the equivalence classes from Definition 3.1, as determined by the profiling hyperplanes. The image also shows a representative point from each equivalence class.

Figure 4: An illustration of the profiling procedure, starting with the initial arrangement in Figure 4(a), adding the profiling hyperplanes in Figure 4(b) and (c), and ending with the equivalence classes from Definition 3.1 illustrated with all representative points in Figure 4(d).

matrix. A face of $\mathcal{V}$ is any nonempty set $\{\boldsymbol{v} \in \mathcal{V} : \tilde{\boldsymbol{B}}\boldsymbol{v} = \boldsymbol{0}\}$ where $\tilde{\boldsymbol{B}}$ is a matrix whose rows are contained in the rows of $\boldsymbol{B}$. The dimension of the face is $k - \text{rank}(\tilde{\boldsymbol{B}})$.

Definition 3.3 (Supporting Hyperplane)**.** *A supporting hyperplane for a convex set $\mathcal{V} \subseteq \mathbb{R}^d$ is a hyperplane with the following properties: (i) $\mathcal{V}$ is contained in one of the closed half-spaces defined by the hyperplane; (ii) $\mathcal{V}$ has at least one boundary point on the hyperplane.*

Now let $\boldsymbol{A}$ be an $n \times d$ matrix, with $d = d_u + d_\theta$, and let $\mathscr{A} := \{\{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{a}_i, \boldsymbol{v} \rangle = 0\} : i = 1, \ldots, n\}$ denote

the linear hyperplane arrangement corresponding to the rows of $\boldsymbol{A}$, where $\boldsymbol{v}$ is a $d \times 1$ vector partitioned as $\boldsymbol{v}^\top = (\boldsymbol{u}^\top, \theta^\top)$.

Definition 3.4 (Profiling Hyperplane)**.** *Let $\mathcal{C}$ be a cell induced by the linear arrangement $\mathscr{A}$, and consider a hyperplane of the form $H = \{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{v} \rangle = 0\}$. Then H is a profiling hyperplane for $\mathcal{C}$ if (i) H is a supporting hyperplane of $\mathcal{C}$, (ii) H is orthogonal to $\mathbb{R}^{d_\theta}$, (iii) the first non-zero element of the normal vector $\boldsymbol{a}$ is 1, (iv) H contains a face of $\mathcal{C}$ defined by the $\tilde{n} \times d$ matrix $\tilde{\boldsymbol{A}} := [\tilde{\boldsymbol{A}}_u \ \tilde{\boldsymbol{A}}_\theta]$ whose rows are contained in $\boldsymbol{A}$, where $\tilde{\boldsymbol{A}}$ and the $\tilde{n} \times d_u$ submatrix $\tilde{\boldsymbol{A}}_u$ satisfy $1 \leq rank(\tilde{\boldsymbol{A}}) \leq d_u + 1$ and $rank(\tilde{\boldsymbol{A}}) - rank(\tilde{\boldsymbol{A}}_u) = 1$.*

Recall that, to recover the equivalence class $\mathcal{S}(\theta)$, it suffices to recover the hyperplanes in $\mathbb{R}^{d_\theta}$ that define the boundaries of $\text{Proj}_\Theta(\mathcal{C})$ (see Proposition 3.2). The boundaries of each projected cell can be characterized by hyperplanes of dimension $d_\theta - 1$, illustrated in Figure 4(d) for the arrangement in Figure 4(a). The following proposition connects the boundaries of the projected cells to profiling hyperplanes.

Proposition 3.3. *Consider the linear arrangement $\mathscr{A}$, let $\mathcal{C}$ be a cell induced by the arrangement, and let $\mathscr{E}(\mathcal{C}) := \{E_1, \ldots, E_M\}$ be the hyperplanes in $\mathbb{R}^{d_\theta}$ that define the unique minimal half-space representation of $Proj_\Theta(\mathcal{C})$.*[33] *Then $E \in \mathscr{E}(\mathcal{C})$ if and only if there exists a profiling hyperplane H for $\mathcal{C}$ such that $E = Proj_\Theta(H)$.*

For example, the profiling hyperplanes associated with the hyperplane arrangement in Figure 4(a) are illustrated in red in Figure 4(b) and Figure 4(c). Note that the $d_\theta - 1$ dimensional hyperplanes defining the boundaries of the projected cells are contained within a unique $d - 1$ dimensional profiling hyperplane (compare Figures 4(c) and 4(d)). To characterize the boundaries of each projected cell, Algorithm 3 aims to return all the profiling hyperplanes associated with the linear arrangement $\mathscr{A}$. By Proposition 3.3, dropping the first d_u elements of the normal vectors for the profiling hyperplanes yields the normal vectors for the hyperplanes in $\mathbb{R}^{d_\theta}$ that define the boundaries of the projected cells. These hyperplanes define an arrangement in $\mathbb{R}^{d_\theta}$, and a final application of our cell enumeration algorithm yields the representative points from each equivalence class.

The definition of a profiling hyperplane is designed to ensure Proposition 3.3 holds. In particular, conditions (i) and (ii) in Definition 3.4 ensure the profiling hyperplanes characterize the boundaries of $\text{Proj}_\Theta(\mathcal{C})$. Condition (iii) is a normalization on the coefficients of the normal vector of each profiling hyperplane. Condition (iv) then ensures that each profiling hyperplane contains only faces of $\mathcal{C}$ whose projection on $\mathbb{R}^{d_\theta}$ is of dimension $d_\theta - 1$.[34] Combined, these conditions ensure the profiling hyperplanes contain the $d_\theta - 1$-dimensional faces of $\text{Proj}_\Theta(\mathcal{C})$, providing a unique representation of boundaries of the projected cell.

Algorithm 3 accepts the normal vectors of a hyperplane arrangement in $\mathbb{R}^d$ as input, and outputs the normal vectors for all profiling hyperplanes. For $1 \leq j \leq d_u + 1$, Algorithm 3 operates by taking size$-j$

[33] That is, the minimal collection of hyperplanes $E_1, \ldots, E_M$ such that $\text{Proj}_\Theta(\mathcal{C})$ can be written as the intersection of open half-spaces whose boundaries are defined by $E_1, \ldots, E_M$.

[34] To see why, let F be a face of $\mathcal{C}$, and let $\Theta(F)$ denote its projection on $\mathbb{R}^{d_\theta}$. By Theorem 2.4 of Balas and Oosten (1998) we have:

$$\dim(\Theta(F)) = \dim(F) - d_u + \text{rank}(\tilde{\boldsymbol{A}}_u) = d_\theta - \text{rank}(\tilde{\boldsymbol{A}}) + \text{rank}(\tilde{\boldsymbol{A}}_u),$$

so that $\dim(\Theta(F)) = d_\theta - 1$ if and only if $\text{rank}(\tilde{\boldsymbol{A}}) - \text{rank}(\tilde{\boldsymbol{A}}_u) = 1$.

Algorithm 3: (Profiling Algorithm)

input : An $n \times d$ matrix $\boldsymbol{A}$, with $d = d_u + d_\theta$ and $n, d_u, d_\theta \geq 1$ (with the last d_θ columns of $\boldsymbol{A}$ corresponding to θ).

output: A matrix $\boldsymbol{V}$ with d_θ columns, characterizing all projected cell boundaries.

1. If any rows in $\boldsymbol{A}$ contain at least d_u leading zeros, include them as rows in $\boldsymbol{V}$. Discard those rows from $\boldsymbol{A}$ and call the remaining matrix $\boldsymbol{A}^*$.
2. **If** $\text{nrow}(\boldsymbol{A}^*) \leq 1$, terminate and return $\boldsymbol{V}$; **else**: let $\mathcal{R}_2$ be the collection of all size-2 subsets of rows from $\boldsymbol{A}^*$.
   (a) Combine each pair of rows into a $2 \times d$ matrix, and convert the matrix to reduced row echelon form.
   (b) Check the rows of the matrix from 2(a) for vectors containing at least d_u leading zeros, and store these vectors in $\boldsymbol{V}$.
   (c) Store each pair of rows yielding a vector with at least d_u leading zeros in $\mathcal{R}_2^*$.
3. For $j = 3, \ldots, d_u + 1$,
   (a) **If** $\text{nrow}(\boldsymbol{A}^*) \leq j - 1$, terminate and return $\boldsymbol{V}$; **else**: let $\mathcal{R}_j$ be the collection of all size-j subsets of rows in matrix $\boldsymbol{A}^*$.
   (b) Discard the subset $R \in \mathcal{R}_j$ if $R^* \subset R$ for any $R^* \in \mathcal{R}_k^*$ with $k < j$.
   (c) **If** no subsets remain in $\mathcal{R}_j$ after step (b), set $\mathcal{R}_j^* = \varnothing$ and move to the next iteration; **else**:
      i. For all surviving subsets $R \in \mathcal{R}_j$, combine the elements of R into a $j \times d$ matrix, and convert the matrix to reduced row echelon form.
      ii. Check the rows of the matrix from step 3(c)i. for vectors with at least d_u leading zeros, and store these vectors in $\boldsymbol{V}$.
      iii. Store those size-j subsets yielding a vector with at least d_u leading zeros in $\mathcal{R}_j^*$, and proceed to the next iteration.

subsets of rows in the normal vector matrix $\boldsymbol{A}$ and performs Gaussian elimination (row reduction) to find profiling hyperplanes, if they exist. Gaussian elimination on size$-j$ subsets of rows will produce a $j \times d$ matrix in reduced row echelon form (RREF). Every non-zero row of the RREF matrix contains a leading 1. Furthermore, any non-zero row of the RREF matrix that contain at least d_u leading zeros defines a linear subspace of $\mathbb{R}^d$ that (i) contains the intersection of j hyperplanes from the arrangement, and (ii) is orthogonal to $\mathbb{R}^{d_u}$. Algorithm 3 stores all these non-zero rows of the RREF matrix that contain at least d_u leading zeros. The following proposition shows that the rows of the matrix returned by Algorithm 3 are exactly the profiling hyperplanes associated with the input arrangement.

Proposition 3.4. *For a cell $\mathcal{C}$ induced by the arrangement $\mathscr{A}$, let $\boldsymbol{A}(\mathcal{C})$ be the matrix whose rows are normals of all profiling hyperplanes of $\mathcal{C}$. Then for any given cell $\mathcal{C}$ induced by the arrangement $\mathscr{A}$ the rows of $\boldsymbol{A}(\mathcal{C})$ are contained in the rows of $\boldsymbol{V}$ returned by Algorithm 3. Furthermore, each row of $\boldsymbol{V}$ returned by Algorithm 3 belongs to a matrix $\boldsymbol{A}(\mathcal{C})$ for some $\mathcal{C}$.*

Proposition 3.4 verifies that our Gaussian elimination algorithm will return all profiling hyperplanes as output. We can also establish an upper bound for the number of rows in the output matrix $\boldsymbol{V}$ from Algorithm

3. We focus on the case when $n > d_u + 1$, the hyperplanes in $\mathscr{A}$ are in general position, and none of them are orthogonal to $\mathbb{R}^{d_\theta}$. For the next result, this corresponds to the worst case.

Proposition 3.5. *Suppose that $n > d_u + 1$, that the hyperplanes in $\mathscr{A}$ are in general position, and that none of the hyperplanes are orthogonal to $\mathbb{R}^{d_\theta}$ (the structural parameter space). Then the number of rows in $\boldsymbol{V}$ returned by Algorithm 3 is $\binom{n}{d_u+1}$.*

For degenerate arrangements, $\boldsymbol{V}$ may be significantly smaller; for instance, when all hyperplanes in $\mathscr{A}$ are orthogonal to $\mathbb{R}^{d_\theta}$, Algorithm 3 terminates after step 1 without any Gaussian elimination operations, in which case $\boldsymbol{V}$ will contain n rows.

The following works through an application of the profiling algorithm to a two player game example, illustrating a case with only two representative points.

Example 2 (Cont'd)**.** *Consider again the complete information game example with two players and pure strategy Nash equilibria, but now consider the following system of equations:*

$$Y_{i1} = \mathbb{1}\{Y_{i2}\theta - U_{i1} \geq 0\},$$
$$Y_{i2} = \mathbb{1}\{Y_{i1}\theta - U_{i2} \geq 0\}.$$

Here θ represents a fixed strategic interaction parameter shared by both players. The hyperplanes for profiling in the example are given by:

$$H_1 = \{(u_1, u_2, \theta) \in \mathbb{R}^3 : -u_1 + \theta = 0\}, \qquad H_2 = \{(u_1, u_2, \theta) \in \mathbb{R}^3 : -u_2 + \theta = 0\},$$
$$H_3 = \{(u_1, u_2, \theta) \in \mathbb{R}^3 : -u_1 = 0\}, \qquad H_4 = \{(u_1, u_2, \theta) \in \mathbb{R}^3 : -u_2 = 0\}.$$

Here $d_u = 2$, $d_\theta = 1$ and $n = 4$, and the matrix $\boldsymbol{A}$ input to Algorithm 3 is given by:

$$\boldsymbol{A} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}.$$

Since none of the rows in $\boldsymbol{A}$ contain a vector with at least d_u leading zeros, no rows will be removed from $\boldsymbol{A}$ in step 1 of Algorithm 3, and $\boldsymbol{A}^* = \boldsymbol{A}$ is sent to step 2. In step 2, there are six $2 \times d$ matrices that can be formed by the size-2 subsets of rows from $\boldsymbol{A}^*$, given by:

$$\mathcal{R}_2 = \{\{\boldsymbol{a}_1, \boldsymbol{a}_2\}, \{\boldsymbol{a}_1, \boldsymbol{a}_3\}, \{\boldsymbol{a}_1, \boldsymbol{a}_4\}, \{\boldsymbol{a}_2, \boldsymbol{a}_3\}, \{\boldsymbol{a}_2, \boldsymbol{a}_4\}, \{\boldsymbol{a}_3, \boldsymbol{a}_4\}\}.$$

In step 2(a) these rows are combined into a matrix and converted to reduced row echelon form:

$$\begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_2^\top \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_3^\top \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_4^\top \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{a}_2^\top \\ \boldsymbol{a}_3^\top \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{a}_2^\top \\ \boldsymbol{a}_4^\top \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ 0 & -1 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{a}_3^\top \\ \boldsymbol{a}_4^\top \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \xrightarrow{RREF} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

Here the pairs $\{\boldsymbol{a}_1, \boldsymbol{a}_3\}$ *and* $\{\boldsymbol{a}_2, \boldsymbol{a}_4\}$ *both yield matrices in reduced row echelon form that contain the vector* $[0\ 0\ 1]$ *with* $d_u = 2$ *leading zeros. The vector is stored in the matrix* $\boldsymbol{V}$*, and at the end of step 2 we have* $\mathcal{R}_2^* = \{\{\boldsymbol{a}_1, \boldsymbol{a}_3\}, \{\boldsymbol{a}_2, \boldsymbol{a}_4\}\}$*. Step 3 starts with* $j = 3$*. There are four size-3 subsets of rows of* $\boldsymbol{A}^*$*, given by:*

$$\mathcal{R}_3 = \{\{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3\}, \{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_4\}, \{\boldsymbol{a}_1, \boldsymbol{a}_3, \boldsymbol{a}_4\}, \{\boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4\}\}.$$

In step 3(b) all of these sets will be discarded, since they each contain at least one subset from $\mathcal{R}_2^*$*. Since no subsets in* $\mathcal{R}_3$ *survive elimination, step 3(c) sets* $\mathcal{R}_3^* = \varnothing$ *and moves to the next iteration. Since* $j = 4$ *exceeds* $d_u + 1 = 3$ *the algorithm terminates and outputs the matrix* $\boldsymbol{V}$ *which contains the single profiling hyperplane with normal vector* $[0\ 0\ 1]$*. The profiling hyperplane divides* $\mathbb{R}^{d_\theta}$ *into two half-spaces. Cell enumeration then yields two profiling points for the example, one with* $\theta > 0$ *and one with* $\theta < 0$*; that is, the linear programs in Theorem 2.2 need only be solved at two values of* θ *in the example.*

Remark 3.1. *From Remark 2.3, the identified set for structural parameters is given by all* $\theta \in \Theta$ *leading to a feasible linear program in Theorem 2.2. In this section, we have discussed conditions under which these linear programs need only be solved at a finite number of representative points for the purpose of bounding counterfactual parameters. Each representative point "represents" an equivalence class, defined by a cone in* Θ*. If a linear program is feasible for a given representative point, then it is feasible for all values of* $\theta \in \Theta$ *belonging to the cone represented by* θ*. Following this argument, it is straightforward to see that the identified set of structural parameters is given by a disjoint union of cones contained in* Θ*.*

Before concluding this section, there is a final gap in the profiling procedure that should be addressed. Recall from Proposition 3.2 that it is impossible to "cross" into a new equivalence class without also either "entering" or "leaving" the projection of some cell, and vice versa. However, Proposition 3.2 hides a subtle point: the projected cells are all open sets, so it is possible to "leave" a projected cell without "entering" a new projected cell.[35] In this case, invoking Proposition 3.2, we must have entered a new equivalence class that lies on boundary between two adjacent projected cells. Our algorithm will not return representative points for these equivalence classes. However, each of these equivalence classes must have an associated collection $\mathcal{S}(\theta)$ completely contained in the collection $\mathcal{S}(\theta')$ for some adjacent projected cell. This is verified in the proof of the following proposition.

[35] For instance, consider the common boundary between two adjacent, open projected cells.

Proposition 3.6. *Consider the arrangement $\mathscr{A}$, and suppose that $\theta \in \Theta$ is such that $\theta \in \partial Proj_\Theta(\mathcal{C}^*)$ for some cell $\mathcal{C}^*$ induced by the arrangement. Then there exists a $\theta' \in Proj_\Theta(\mathcal{C}^*)$ such that $\mathcal{S}(\theta) \subset \mathcal{S}(\theta')$.*

Combined with the second part of Proposition 3.1 (and the discussion following it), this result implies that the boundaries of the projected cells can be safely ignored when selecting representative points in our profiling procedure.

3.3 Stitching and the Minimal Relevant Partition

The output of the cell enumeration step (for a given representative point) is a set of witness points and a set of sign vectors characterizing all cells formed by the arrangement of hyperplanes $\mathscr{H}$. Although Theorem 2.1 suggests the cells induced by the arrangement $\mathscr{H}$ are sufficient for the purposes of bounding counterfactuals, not all cells are necessary. In particular, given a counterfactual γ, there may exist disjoint cells $C_1, C_2 \in \mathscr{C}$ such that for every $\boldsymbol{u} \in C_1$ and $\boldsymbol{u}' \in C_2$ we have $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta)$ for all $\boldsymbol{x} \in \mathcal{X}$. A simple case can be seen in Figure 1(b) for the game example, which illustrates multiple cells that produce the same $\boldsymbol{Y}-$level sets. Any such cells can be combined together, or "stitched," without affecting the results of Theorem 2.1 or Theorem 2.2. Stitching can be used to dramatically reduce the number of parameters and constraints in the optimization problems of Theorem 2.2, leading to further computational improvements.[36] Stitching relies only on the output of our cell enumeration procedure, and so is applicable even when additional assumptions are imposed, such as those introduced in the next section.

Stitching produces a minimal collection of sets in the latent variable space. In a multinomial choice model (e.g. Example 1) with quasi-linear utility stitching can be used to construct the minimal relevant partition (MRP) of Tebaldi et al. (2021).[37] Stitching can also be used to produce the MRP in a considerably more general class of models. To make the claim rigorous, let $\mathcal{H}(\theta)$ denote the union of all hyperplanes of the form:

$$\left\{\boldsymbol{u} \in \mathcal{U} : g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\right\}, \tag{3.6}$$

from Assumption 2.1. Now consider the following sets:

$$\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta) := \text{int}\left\{\boldsymbol{u} \in \mathcal{U} \setminus \mathcal{H}(\theta) : \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\right\}, \tag{3.7}$$

$$\mathcal{U}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \theta) := \text{int}\left\{\boldsymbol{u} \in \mathcal{U} \setminus \mathcal{H}(\theta) : \boldsymbol{y} \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\right\}. \tag{3.8}$$

We now present a generalized version of the definition of the MRP in Tebaldi et al. (2021) that applies to the broader class of models and counterfactuals in this paper.

[36] Similar considerations arise when using identification methods based on Artstein's inequalities, in which case focus has been on finding "core-determining classes" (c.f. Galichon and Henry (2011), Chesher and Rosen (2017)). Analogous to the minimal relevant partition, a small core-determining class has computational benefits. It is still not known how to find the smallest core-determining class for a general class of models.

[37] Tebaldi et al. (2021) also find that the MRP delivers a much smaller collection of sets than the core-determining class of Chesher et al. (2013), which can lead to substantial computational benefits in large examples.

Definition 3.5 (MRP, Generalized)**.** *Consider an environment where Assumptions 2.1 and 2.2 hold. Given a counterfactual γ the minimal relevant partition (MRP) of the latent space $\mathcal{U}$ is the collection $\mathbb{U}$ of sets $\mathcal{U}' \subset \mathcal{U}$ with the following properties:*

(i) The sets in $\mathbb{U}$ form a partition of the set:

$$\left(\bigcup_{(\boldsymbol{y},\boldsymbol{x})} \mathcal{U}(\boldsymbol{y},\boldsymbol{x},\theta) \right) \cup \left(\bigcup_{(\boldsymbol{y},\boldsymbol{x})} \mathcal{U}_\gamma(\boldsymbol{y},\boldsymbol{x},\theta) \right). \tag{3.9}$$

(ii) Any two elements $\boldsymbol{u}$ and $\boldsymbol{u}'$ from (3.9) belong to the same set $\mathcal{U}' \in \mathbb{U}$ if and only if $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta) = \mathcal{Y}(\boldsymbol{x},\boldsymbol{u}',\theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta) = \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}',\theta)$ for all $\boldsymbol{x} \in \mathcal{X}$.

The definition differs from the MRP definition in Tebaldi et al. (2021) in a few key ways. First, the minimal relevant partition is a partition of (3.9), and not $\mathcal{U}$. In our general class of models, there may be large regions of $\mathcal{U}$ that cannot rationalize any factual or counterfactual outcome because of model incoherency. To construct the *minimal* relevant partition of the latent variable space, these regions should be excluded. This issue does not arise in the model of Tebaldi et al. (2021). Second, our MRP depends on the counterfactual outcome sets $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)$ through condition (ii). Intuitively, this is because counterfactuals that modify endogenous outcome variables also modify the underlying system of structural equations, which can change the relevant partition of the latent variable space. Since Tebaldi et al. (2021) do not consider counterfactuals that modify endogenous outcome variables the consideration does not arise in their context. Third, the MRP of Definition 3.5 is unique, while the MRP proposed by Tebaldi et al. (2021) is not unique.[38] This arises because of our slightly different approach to deal with the cell boundaries, or hyperplanes. Both this paper and the paper of Tebaldi et al. (2021) assume that zero probability is assigned to the hyperplanes by the latent variable distribution. The MRP in Tebaldi et al. (2021) is defined to be unique up to a Lebesgue null set. In our approach, the hyperplanes are subtracted from all sets in the MRP.[39] Uniqueness of the MRP in our approach is formally established in Lemma D.1 in Appendix D.

The following result asserts that the MRP can be constructed by taking a disjoint union of cells induced by the arrangement $\mathscr{H}$.

Lemma 3.1 (Stitching Lemma)**.** *Suppose Assumption 2.1 and 2.2 hold, and let $\mathbb{U}$ denote the MRP for a given counterfactual γ. Then every $\mathcal{U}' \in \mathbb{U}$ is the disjoint union of the cells $\mathscr{C}$ induced by the arrangement $\mathscr{H}$. Furthermore, two cells $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}$ are contained in the same set $\mathcal{U}' \in \mathbb{U}$ if and only if their witness points $\boldsymbol{u}_1 \in \mathcal{C}_1$ and $\boldsymbol{u}_2 \in \mathcal{C}_2$ satisfy $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u}_1,\theta) = \mathcal{Y}(\boldsymbol{x},\boldsymbol{u}_2,\theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_1,\theta) = \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_2,\theta)$ for all $\boldsymbol{x} \in \mathcal{X}$.*

Lemma 3.1 states that construction of the MRP requires only the output of the cell enumeration procedure. In particular, the algorithmic process of stitching takes the sign vectors from the cell enumeration

[38] See Appendix S3 in Tebaldi et al. (2021). Depending on how the points on the boundaries are assigned to the sets in their MRP, it is possible to construct an uncountably infinite number of (Lebesgue almost-everywhere equivalent) MRPs.

[39] In particular, the elements of the latent variable space lying exactly on the hyperplanes are subtracted when defining the sets (3.7) and (3.8).

step as input, and stitches cells by determining which cells induce the same set of factual and counterfactual outcomes for every $\boldsymbol{x} \in \mathcal{X}$. We refer to Tebaldi et al. (2021) Appendix S5 for a detailed comparison of the MRP to approaches based on Artstein's inequalities and core-determining classes. We also present some results relating the MRP to core-determining classes for the identification of structural parameters in Appendix D. We find that the MRP is not a core-determining class when the model is incomplete and incoherent, which serves to further distance the two approaches.[40] On the other hand, we find that the MRP is a core-determining class for certain counterfactuals when the model is complete and coherent; this is the case, for instance, in the model of Tebaldi et al. (2021).

4 Additional Assumptions

This section introduces additional assumptions that can be used to tighten the bounds on counterfactual parameters that arise as a result of using our method. Whether an assumption can be accommodated by our framework is determined by whether a version of Theorem 2.1 holds when the additional assumption is imposed. Here we focus on explaining the assumptions that are used in the application section.

4.1 Independence Assumptions

In many environments the researcher has access to random variables that are plausibly independent of the vector latent variables. These variables may enter the model in two possible ways: they may enter non-trivially in one or more of the structural functions $\varphi_1^{(m_1,m_2)}, \ldots, \varphi_{d_y}^{(m_1,m_2)}$ from Assumption 2.1, or they may be excluded from all structural equations. In either case, these independent variables can provide identifying information by inducing variation in the observed distribution without affecting the distribution of latent variables.

Assumption 4.1 (Independence)**.** *There exists a subset $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ such that $P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}) = P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}')$ for all $B \in \mathfrak{B}(\mathcal{U})$ and all $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ occurring with positive probability.*

Assumption 4.1 can be used to impose independence of the vector $\boldsymbol{U}$ from a subvector of $\boldsymbol{X}$, possibly conditional on some other subvector of $\boldsymbol{X}$, by properly choosing the set $\widetilde{\mathcal{X}}$. Additional details are presented in Appendix B.1 on how to extend the main definitions and results of the previous sections to the case when Assumption 4.1 holds. The proofs in Appendix B.1 reveal that the assumption can be imposed on the vector $\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}$ from Theorem 2.2 by simply adding constraints of the form:

$$\sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\boldsymbol{y} \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) = \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\boldsymbol{y} \in \mathcal{Y}} \pi_{\boldsymbol{x}'}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell),$$

for every $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$, and for $\ell = 1, \ldots, L$. These constraints are linear in the parameter vector $\boldsymbol{\pi}$, and so

[40] This does not contradict any results in this paper; instead, it suggests that naively using the MRP as test sets for Artstein's inequalities in an incomplete or incoherent model may discard relevant identifying information.

the optimization problems in Theorem 2.2 remain as linear programs when independence restrictions are imposed.

4.2 Quantile and Sign Restrictions

In addition to independence restrictions, we can impose median zero and sign restrictions on elements of $\boldsymbol{U}$.

Assumption 4.2 (Quantile and Sign Restrictions)**.** *There exists elements $U_{k_1}, \ldots, U_{k_Q}$ of the vector $\boldsymbol{U}$ with supports $\mathcal{U}_{k_1}, \ldots, \mathcal{U}_{k_Q}$ such that:*

$$P_{U_{k_q}}(U_{k_q} \leq \tau_{k_q}) = p_{k_q}, \tag{4.1}$$

for known pairs (p_{k_q}, τ_{k_q}), for $q = 1, \ldots, Q$.

Assumption 4.2 permits us to impose assumptions on the quantiles of various elements of the vector $\boldsymbol{U}$. For instance, taking $(p_{k_q}, \tau_{k_q}) = (0.5, 0)$ in Assumption 4.2 imposes a median zero restriction on the element U_{k_q}. In addition, taking $(p_{k_q}, \tau_{k_q}) = (1, 0)$ imposes a sign restriction. The assumption can be combined with Assumptions 3.1 and Assumption 4.1 in any way, and can also be implemented conditional on $\boldsymbol{X}$.

The partition of the latent variable space introduced before Theorem 2.1 must be refined to impose quantile restrictions of the form (4.1). Define $\boldsymbol{e}_{k_q}$ as the basis vector with 1 as the k_q^{th} element and zeroes elsewhere. Now consider the collection of hyperplanes of the form:

$$E_{k_q} := \left\{ \boldsymbol{u} \in \mathbb{R}^{d_u} : \langle \boldsymbol{e}_{k_q}, \boldsymbol{u} \rangle = \tau_{k_q} \right\}.$$

These additional hyperplanes must be added to the arrangement $\mathscr{H}$ before using our procedure since the hyperplanes E_{k_q} can intersect the interior of the cells formed by the initial arrangement. To impose the quantile assumption, these cells must be split into separate cells. Additional details are presented in Appendix B.1 on how to extend the main definitions and theorems of the previous sections to the case when Assumption 4.1 holds.

4.3 Counterfactual Equilibrium Selection Constraints

To describe our counterfactual equilibrium selection constraints, we begin by introducing some additional terminology.

Definition 4.1 (Trivial Counterfactual)**.** *A counterfactual γ is trivial at $\boldsymbol{x} \in \mathcal{X}$ if: (i) $c_x(\boldsymbol{x}) = \boldsymbol{x}$, (ii) $\Pi^\star_{(m_1,m_2)} = \varnothing$ for all $m_1 \in \mathcal{M}_1$ and $m_2 \in \mathcal{M}_2$, (iii) $\overline{\mathcal{Y}}^\star_{(m_1,m_2)} = \overline{\mathcal{Y}}_{(m_1,m_2)}$ for all $m_1 \in \mathcal{M}_1$ and $m_2 \in \mathcal{M}_2$, (iv) $\mathcal{M}_1^\star = \mathcal{M}_1$ and $\mathcal{M}_2 = \mathcal{M}_2^\star$.*

Intuitively, a counterfactual is trivial at $\boldsymbol{x} \in \mathcal{X}$ if executing the counterfactual requires no action; it represents the "do nothing" counterfactual. The trivial counterfactual highlights an issue in incomplete models. Most approaches to modeling counterfactuals implicitly assume a notion of reproducibility: for

a given pair $(\boldsymbol{y}, \boldsymbol{x})$, we expect to observe $\boldsymbol{y}$ (the factual outcome) after a trivial counterfactual has been performed.[41] For our class of counterfactuals the idea of reproducibility is satisfied automatically in complete models, since for any trivial counterfactual the values of both $\boldsymbol{Y}$ and $\boldsymbol{Y}_\gamma$ are completely determined by the same (unchanged) system of equations. However, notions of reproducibility can fail in models that are incomplete since the nonlinear system of equations determining $\boldsymbol{Y}$ and $\boldsymbol{Y}_\gamma$ can admit multiple solutions. The solution "selected" from the set of solutions in the system determining $\boldsymbol{Y}$ may be different from the solution "selected" from the set of solutions to the system determining $\boldsymbol{Y}_\gamma$, even if these two systems are identical. This arises when the counterfactual equilibrium selection mechanism is different from the factual equilibrium selection mechanism.

The following assumption imposes equivalence between the counterfactual and factual equilibrium selection mechanisms for trivial counterfactuals, which can be used to further narrow the bounds on counterfactual outcomes.[42]

Assumption 4.3. *For any $\boldsymbol{x} \in \mathcal{X}$ such that γ is trivial at $\boldsymbol{x} \in \mathcal{X}$, we have:*

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right), \tag{4.2}$$

$P_{\boldsymbol{U},\boldsymbol{X}}-a.s$ for all $\boldsymbol{y} \in \mathcal{Y}$.

Note that when the model is complete, (4.2) holds vacuously for the trivial counterfactuals under consideration. In this sense, a version of Assumption 4.3 holds automatically when the model is complete. When the model is incomplete the probability on the left side of (4.2) is exactly the counterfactual equilibrium selection mechanism, and the probability on the right is the factual equilibrium selection mechanism. Equating the two imposes exactly that the same solution is selected from the nonlinear system determining $\boldsymbol{Y}$ and $\boldsymbol{Y}_\gamma$ under any trivial counterfactual.

5 Application

In this section we consider two illustrations of our procedure: a mobile phone multinomial choice example and an airline entry game example. All bounds on counterfactual functionals in this section are computed using a modified plug-in estimator proposed by Gu and Russell (2021), which is consistent under weak assumptions. The estimator involves a small but vanishing relaxation of the linear program constraints, and its consistency is discussed in detail in Appendix E. There are also a number of recent papers on hypothesis testing and confidence set construction in settings similar to ours (e.g. Deb et al. (2018), Kitamura and Stoye (2018), Gafarov (2019), Fang et al. (2020), Cho and Russell (2022)). Because of the very high-dimensional nature of the our linear programming problems (constraints and parameters), and since both our objective function

[41] For instance, this is implicit in individual-level counterfactuals in a potential outcome model.

[42] As far as we are aware the only other paper with a similar discussion of factual and counterfactual equilibrium selection mechanisms is Canen and Song (2020) in the context of empirical games.

and constraints can have data-dependent gradients, existing methods are either not applicable, extremely conservative, or are computationally infeasible. We leave the development of suitable inference methods for future work.

5.1 Keane and Wasi (2013) Mobile Phone Choice

Modeling demand for differentiated products is a central goal of empirical industrial organization and marketing research. Here we consider an application to demand for mobile phones, allowing for the possibility that individual choice sets are unobserved. Our application is based on a sample from the Center for the Study of Choice at the University of Technology Sydney, and was previously analyzed in Keane and Wasi (2013). The data were collected in an experimental setting where three phones with varying characteristics and prices were offered to the participants. Each participant either chose one of the three phones, or chose no phone. A total of 3944 choices were observed. We assume that consumer choices were made according to the following model:

$$Y_i = \arg\max_{j \in \mathcal{J}_i} \xi_{ij}, \quad \xi_{ij} = P_{ij}\beta_i^P + W_j\beta_i^W + U_{ij}, \quad j = 0, 1, 2, 3,$$

where ξ_{ij} represents the utility of individual i from choosing phone j, P_{ij} denotes the price faced by consumer i for phone j (in hundreds of Australian dollars (AUD)), W_j is an indicator for whether phone j has WiFi capability, $\mathcal{J}_i$ is the choice set faced by individual i, and $j = 0$ represents the outside option of choosing no phone. We discretize the price of each phone in increments of 50 AUD, with a total of 11 possible prices. We set both the price and WiFi variable to zero for the outside option. To capture taste heterogeneity for the observed characteristics, we include specifications where the coefficients on price and WiFi are random and individual-specific. Since only utility differences matter, in the most flexible model there are five latent variables contained in the vector $\boldsymbol{U}_i := (\beta_i^P, \beta_i^W, \Delta U_{i1}, \Delta U_{i2}, \Delta U_{i3})$ with $\Delta U_{ij} = U_{ij} - U_{i0}$.

We consider two models and two counterfactuals. The first model is a fixed choice set model where $\mathcal{J}_i = \{0, 1, 2, 3\}$ for $i = 1, \ldots, n$, and the second model is a heterogeneous choice set model where $\mathcal{J}_i \in \{\{0, 1, 2, 3\}, \{0, 1, 2\}, \{0, 1, 3\}, \{0, 2, 3\}\}$ for $i = 1, \ldots, n$. The first counterfactual is for the fixed choice set model, and considers the average change in the probability of choosing phone 1 when the price of phone 1 is reduced from 250 AUD to 50 AUD, and the prices of phone 2 and 3 are kept fixed at 250 AUD and 150 AUD. The second counterfactual is for the heterogeneous choice set model, and considers the average change in the probability of choosing phone 1 when the choice set is changed from $\mathcal{J}_i = \{0, 1, 2, 3\}$ $\forall i$ to $\mathcal{J}_i = \{0, 1, 2\}$ $\forall i$. These counterfactuals are summarized in Table 3 using the notation in the paper. For both models, we investigate how the bounds change under various combinations of the following assumptions:

(I1) Partial independence: $\boldsymbol{U}_i \perp W_j$.

(I2) Full independence: $\boldsymbol{U}_i \perp (P_{ij}, W_j)$.

(MZ) A median zero restriction: $\text{Median}(\Delta U_{ij}) = 0$ for $j = 1, 2, 3$.

Label	Objective	Description	$\mathcal{M}_2$	γ_1		γ_2	
				c_x	$\mathcal{M}_2^\star$	c_x	$\mathcal{M}_2^\star$
(P1)	$P_{\mathbf{Y}_\gamma}(Y_{i\gamma_1}=1)$ $-P_{\mathbf{Y}_\gamma}(Y_{i\gamma_2}=1)$	Effect of reducing the price of phone 1	$\{1\}$	$c_x(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$ $=(0.5,2.5,1.5,W_1,W_2,W_3)$	$\{1\}$	$c_x(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$ $=(2.5,2.5,1.5,W_1,W_2,W_3)$	$\{1\}$
(P2)	$P_{\mathbf{Y}_\gamma}(Y_{i\gamma_1}=1)$ $-P_{\mathbf{Y}_\gamma}(Y_{i\gamma_2}=1)$	Effect of eliminating phone 3 from the choice set	$\{1,2,3,4\}$	$c_x(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$ $=(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$	$\{2\}$	$c_x(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$ $=(P_{i1},P_{i2},P_{i3},W_1,W_2,W_3)$	$\{1\}$

Table 3: A description of the two counterfactuals considered for the phone choice example. Here $\mathcal{M}_2$ is an index set for the choice sets faced by the consumer. We label the choice sets $\{0,1,2,3\}$, $\{0,1,2\}$, $\{0,1,3\}$, $\{0,2,3\}$ with the indices $1,2,3,4$. For instance, $\mathcal{M}_2=\{1\}$ means the choice set is fixed at $\{0,1,2,3\}$.

(FC) Fixed coefficients: $\beta_i^P=\beta^P$ and $\beta_i^W=\beta^W$ for $i=1,\ldots,n$.

(SR1) A sign restriction on the price coefficient: $\beta_i^P\leq 0$.

(SR2) A sign restriction on the WiFi coefficient: $\beta_i^W\geq 0$.

The results for the first counterfactual in the fixed choice set model are displayed in the column labelled *(P1)* in Table 4. Without any restrictions, the bounds are wide, but when the coefficient on price is restricted to be negative, the effect of reducing the price of phone 1 is bounded below by zero. Further restricting the coefficient on WiFi to be positive and adding a median zero restriction tightens the upper bound. The tightest nonempty interval for the counterfactual is $[0.00, 0.48]$, which is obtained when all assumptions except (I2) are imposed. This bound indicates that, without parametric distributional assumptions on the latent variables, we cannot rule out that large price changes may induce small substitution effects. Intuitively, with largely unrestricted consumer preference heterogeneity, consumer preferences for phone 2, phone 3, or the outside option may be so strong that there is a no observed substitution effect when the price of phone 1 is reduced. On the other hand, the upper bound suggests that nearly half of all consumers might be willing to switch their phone choice to phone 1 given the substantial price reduction. We also find that the bounds are empty for the fixed choice set model whenever we impose full independence. Since the bounds are nonempty when imposing partial independence, this result suggests that price may be endogenous. Since price is mechanically independent of latent consumer tastes in the experimental setting in our data, this is likely arising due to the correlation between price and latent product attributes valued by the consumer, which will enter the utility function through the product-specific utility shock.

The results for the second counterfactual in the heterogeneous choice set model are displayed in the column labelled *(P2)* in Table 4. Under all combinations of assumptions the lower bound for the counterfactual effect is zero, indicating that removing phone 3 from the full choice set may have no effect on the proportion of consumers choosing phone 1. This result can be rationalized by an environment where none of the consumers rank phone 1 second to phone 3. Zero must also be the nonparametric lower bound, since removing phone 3 cannot reduce the probability of choosing phone 1. When no assumptions are imposed, the upper bound on the counterfactual effect is 0.29, which is the sum of the observed choice probabilities for phone 1 and

Table 4: Results for the phone choice application. The definition of each counterfactual outcome is provided in Table 3. The "Modal # of Cells" column indicates the modal (most frequent) number of latent space cells found by the cell enumeration algorithm, with the mode taken across all representative points.

(I1)	*(I2)*	*(MZ)*	*(FC)*	*(SR1)*	*(SR2)*	Representative Points	Modal # of Cells	*(P1)*	*(P2)*
						-	1500470	$[-0.70, 0.93]$	$[0.00, 0.29]$
					✓	-	1515968	$[-0.52, 0.87]$	$[0.00, 0.29]$
				✓		-	1500666	$[0.00, 0.93]$	$[0.00, 0.29]$
				✓	✓	-	1516188	$[0.00, 0.87]$	$[0.00, 0.29]$
			✓			208	51711	$[-0.70, 0.91]$	$[0.00, 0.29]$
			✓		✓	104	51711	$[-0.52, 0.86]$	$[0.00, 0.29]$
			✓	✓		104	51711	$[0.00, 0.91]$	$[0.00, 0.29]$
			✓	✓	✓	52	51711	$[0.00, 0.86]$	$[0.00, 0.29]$
		✓				-	1691050	$[-0.70, 0.50]$	$[0.00, 0.29]$
		✓			✓	-	1709084	$[-0.52, 0.50]$	$[0.00, 0.29]$
		✓		✓		-	1691246	$[0.00, 0.49]$	$[0.00, 0.29]$
		✓		✓	✓	-	1709304	$[0.00, 0.49]$	$[0.00, 0.29]$
		✓	✓			208	56521	$[-0.70, 0.49]$	$[0.00, 0.29]$
		✓	✓		✓	104	56521	$[0.00, 0.49]$	$[0.00, 0.29]$
		✓	✓	✓		104	56521	$[0.00, 0.49]$	$[0.00, 0.29]$
		✓	✓	✓	✓	52	56521	$[0.00, 0.49]$	$[0.00, 0.29]$
✓						-	1500470	$[-0.52, 0.85]$	$[0.00, 0.29]$
✓					✓	-	1515968	$[-0.52, 0.85]$	$[0.00, 0.29]$
✓				✓		-	1500666	$[0.00, 0.85]$	$[0.00, 0.29]$
✓				✓	✓	-	1516188	$[0.00, 0.85]$	$[0.00, 0.29]$
✓			✓			208	51711	$[-0.52, 0.85]$	$[0.00, 0.29]$
✓			✓		✓	104	51711	$[-0.52, 0.85]$	$[0.00, 0.29]$
✓			✓	✓		104	51711	$[0.00, 0.85]$	$[0.00, 0.29]$
✓			✓	✓	✓	52	51711	$[0.00, 0.85]$	$[0.00, 0.29]$
✓		✓				-	1691050	$[-0.52, 0.50]$	$[0.00, 0.29]$
✓		✓			✓	-	1709084	$[-0.52, 0.50]$	$[0.00, 0.29]$
✓		✓		✓		-	1691246	$[0.00, 0.48]$	$[0.00, 0.29]$
✓		✓		✓	✓	-	1709304	$[0.00, 0.48]$	$[0.00, 0.29]$
✓		✓	✓			208	56521	$[0.00, 0.48]$	$[0.00, 0.29]$
✓		✓	✓		✓	104	56521	$[0.00, 0.48]$	$[0.00, 0.29]$
✓		✓	✓	✓		104	56521	$[0.00, 0.48]$	$[0.00, 0.29]$
✓		✓	✓	✓	✓	52	56521	$[0.00, 0.48]$	$[0.00, 0.29]$
	✓					-	1500470	∅	$[0.00, 0.25]$
	✓				✓	-	1515968	∅	$[0.00, 0.25]$
	✓			✓		-	1500666	∅	$[0.00, 0.23]$
	✓			✓	✓	-	1516188	∅	$[0.00, 0.23]$
	✓		✓			208	51711	∅	$[0.00, 0.23]$
	✓		✓		✓	104	51711	∅	$[0.00, 0.23]$
	✓		✓	✓		104	51711	∅	$[0.00, 0.23]$
	✓		✓	✓	✓	52	51711	∅	$[0.00, 0.23]$
	✓	✓				-	1691050	∅	$[0.00, 0.25]$
	✓	✓			✓	-	1709084	∅	$[0.00, 0.25]$
	✓	✓		✓		-	1691246	∅	$[0.00, 0.23]$
	✓	✓		✓	✓	-	1709304	∅	$[0.00, 0.22]$
	✓	✓	✓			208	56521	∅	$[0.00, 0.22]$
	✓	✓	✓		✓	104	56521	∅	$[0.00, 0.22]$
	✓	✓	✓	✓		104	56521	∅	$[0.00, 0.22]$
	✓	✓	✓	✓	✓	52	56521	∅	$[0.00, 0.22]$

phone 3. This value is the nonparametric upper bound for the probability of choosing phone 1, and can be rationalized by an environment where (i) everyone who chose phone 1 in the data faced the choice set $\mathcal{J}_i = \{0, 1, 2\}$ and actually preferred phone 3, and (ii) everyone who chose phone 3 in the data ranked phone 1 as their second-preferred option. In this setting, when given the full choice set $\mathcal{J}_i = \{0, 1, 2, 3\}$ no one chooses phone 1, but everyone switches from phone 3 to phone 1 when given the choice set $\mathcal{J}_i = \{0, 1, 2\}$. Unlike the first counterfactual, imposing any combination of (I1), (MZ), (FC), (SR1) or (SR2) has no impact on the bounds, although there is a small impact on the bounds when assumption (I2) is imposed. Also unlike the first counterfactual, the bounds are nonempty for all assumptions.

Finally, for all specifications, the number of representative points is relatively small, but the number of latent space cells is large: over 1.5 million in half of the specifications. Despite the large number of cells, our cell enumeration algorithm is able to enumerate all cells in less than a minute on a standard desktop computer. The major computational bottleneck for this example arises at the linear programming stage where there is a high demand on short-term computer memory in order to store the constraints.

5.2 Kline and Tamer (2016) Airline Entry Game Example

In this section we apply our method to an airline entry game example using data from Kline and Tamer (2016). The data is from the second quarter of the 2010 Airline Origin and Destination Survey, and contains information from $n = 7,882$ markets, defined as origin-destination pairs. As in Kline and Tamer (2016) we consider the entry behaviour of low cost carriers (LCs) and other airlines (OAs). Competition between new LCs and incumbent legacy carriers in the US, and the effects of the competition on ticket prices, has been an important object of study in empirical industrial organization (c.f. Goolsbee and Syverson (2008), Brueckner et al. (2013), Tan (2016)). The LCs include AirTran, Allegiant Air, Frontier, JetBlue, Midwest Air, Southwest, Spirit, Sun Country, USA3000, and Virgin America, and the OAs include all other airlines. Market entry is determined by whether an airline services the origin-destination route for that market. Both the LCs and OAs are treated as a single carrier, and an LC (OA) carrier enters a market if at least one of the LC (OA) carriers services that market.

A market is indexed by $i = 1, \ldots, n$, and Y_i^{LC} and Y_i^{OA} are binary variables representing the entry decisions of the LC and OA carriers. The payoff functions for the two carriers are specified as:

$$\pi^{LC} = \beta_1^{LC} M_i^{LC} + \beta_2^{LC} S_i + \delta_i^{LC} Y_i^{OA} - \varepsilon_i^{LC},$$
$$\pi^{OA} = \beta_1^{OA} M_i^{OA} + \beta_2^{OA} S_i + \delta_i^{OA} Y_i^{LC} - \varepsilon_i^{OA},$$

where M_i^{LA}, M_i^{OA}, and S_i are covariates introduced below, and ε_i^{LC} and ε_i^{OA} are payoff shocks that are unobserved to the econometrician, but observed by both carriers.[43] Carriers enter a market if the payoff from doing so is positive. We assume that the carriers make entry decisions simultaneously and we focus on

[43] Note that our nonparametric treatment of the latent variable distribution implies that including a constant in either payoff function is redundant, since it can be absorbed into ε_i^{LC} and ε_i^{OA}.

pure strategy Nash equilibria. In this setting the entry decisions of the two carriers can be characterized by the following simultaneous discrete choice model:

$$Y_i^{LC} = \mathbb{1}\left\{\beta_1^{LC} M_i^{LC} + \beta_2^{LC} S_i + \delta_i^{LC} Y_i^{OA} \geq \varepsilon_i^{LC}\right\}, \tag{5.1}$$

$$Y_i^{OA} = \mathbb{1}\left\{\beta_1^{OA} M_i^{OA} + \beta_2^{OA} S_i + \delta_i^{OA} Y_i^{LC} \geq \varepsilon_i^{OA}\right\}. \tag{5.2}$$

The variable M_i^j represents the market presence of carrier $j \in \{LC, OA\}$ in market i. A large airport presence for a particular airline is thought to deter entry of other airlines, in addition to being a direct determinant of the airline's market power and profitability (c.f. Borenstein (1989), Berry (1992), Evans and Kessides (1993), Ciliberto and Williams (2010)). To compute these variables, market i is divided into an origin and destination city, and the number of routes serviced by each airline is determined for both the origin and destination cities. The number of routes serviced by carrier j from the origin city is then divided by the total number of routes serviced by any airline from the origin city to provide a measure of market presence of airline j in the origin city for market i. The same procedure is repeated for the destination city. The market presence variable for each airline is computed as the average of the market presence for both the origin and destination cities, and the market presence variable M_i^j for the LC and OA carriers is taken as the maximum of the market presence variables for each of their constituent airlines. We also include a variable S_i measuring the size of market i using population data collected from US Census Bureau. We then define S_i as the maximum of the origin and destination city populations from 2010. Unlike the market presence variable, the market size variable varies across markets but not across carriers. In our implementation the market presence variables are discretized into three values using its 0.33 and 0.67 quantiles, and the market size variable is discretized into two values depending on whether it is above or below the median.

In our baseline model, we allow both strategic interaction parameters δ_i^{LC} and δ_i^{OA} to be random. We also do not impose any restrictions on the joint distribution of $(Y_i^{LC}, Y_i^{OA}, M_i^{LC}, M_i^{OA}, S_i, \delta_i^{LC}, \delta_i^{OA}, \varepsilon_i^{LC}, \varepsilon_i^{OA})$, other than any restrictions implied by the system (5.1) and (5.2). In particular, we allow for arbitrary dependence between ε_i^{LC} and ε_i^{OA}, both within and across markets, and allow for all variables entering the payoff function to be endogenous, in the sense that they can depend nontrivially on the latent variables $(\delta_i^{LC}, \delta_i^{OA}, \varepsilon_i^{LC}, \varepsilon_i^{OA})$. Starting from our baseline model, we then consider the following assumptions:

(I) An independence condition: $(M_i^{LC}, M_i^{OA}) \perp\!\!\!\perp (\delta_i^{LC}, \delta_i^{OA}, \varepsilon_i^{LC}, \varepsilon_i^{OA})$.[44]

(EQ) A counterfactual equilibrium selection restriction: for $\boldsymbol{X}_i = (M_i^{LC}, M_i^{OA}, S_i)$, and for any value $\boldsymbol{x}$ such that γ is trivial, we have:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}((Y_{i\gamma}^{LC}, Y_{i\gamma}^{OA}) = \boldsymbol{y} \mid \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}((Y_i^{LC}, Y_i^{LC}) = \boldsymbol{y} \mid \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{U}_i = \boldsymbol{u}),$$

$P_{\boldsymbol{U},\boldsymbol{X}}$−a.s. for all $\boldsymbol{y} \in \{0,1\}^2$.

(MZ) A median zero restriction: $\text{Median}(\varepsilon_i^{LC}) = \text{Median}(\varepsilon_i^{OA}) = 0$.

[44] Joint independence $(M_i^{LC}, M_i^{OA}, S_i) \perp\!\!\!\perp (\delta_i^{LC}, \delta_i^{OA}, \varepsilon_i^{LC}, \varepsilon_i^{OA})$ leads to empty bounds in all specifications.

Label	Objective	Description	γ_1			γ_2		
			c_x	$y^\star$	$\Pi^\star$	c_x	$y^\star$	$\Pi^\star$
(KT1)	$P_{\boldsymbol{Y}_\gamma}(Y^{LC}_{i\gamma_1}=1)$ $-P_{\boldsymbol{Y}_\gamma}(Y^{LC}_{i\gamma_2}=1)$	ATE on Y_i^{LC} of moving Y_i^{OA} from $0\to 1$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},M_i^{OA},S_i)$	$(\cdot\,,1)$	$\{2\}$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},M_i^{OA},S_i)$	$(\cdot\,,0)$	$\{2\}$
(KT2)	$P_{\boldsymbol{Y}_\gamma}(Y^{OA}_{i\gamma_1}=1)$ $-P_{\boldsymbol{Y}_\gamma}(Y^{OA}_{i\gamma_2}=1)$	ATE on Y_i^{OA} of moving Y_i^{LC} from $0\to 1$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},M_i^{OA},S_i)$	$(1,\cdot\,)$	$\{1\}$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},M_i^{OA},S_i)$	$(0,\cdot\,)$	$\{1\}$
(KT3)	$P_{\boldsymbol{Y}_\gamma}(Y^{LC}_{i\gamma_1}=1)$ $-P_{\boldsymbol{Y}_\gamma}(Y^{LC}_{i\gamma_2}=1)$	ATE on Y_i^{LC} of moving M_i^{OA} from $1\to 3$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},3,S_i)$	–	$\varnothing$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(M_i^{LC},1,S_i)$	–	$\varnothing$
(KT4)	$P_{\boldsymbol{Y}_\gamma}(Y^{OA}_{i\gamma_1}=1)$ $-P_{\boldsymbol{Y}_\gamma}(Y^{OA}_{i\gamma_2}=1)$	ATE on Y_i^{OA} of moving M_i^{LC} from $1\to 3$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(3,M_i^{OA},S_i)$	–	$\varnothing$	$c_x(M_i^{LC},M_i^{OA},S_i)$ $=(1,M_i^{OA},S_i)$	–	$\varnothing$

Table 5: A description of the four counterfactuals considered for the Kline and Tamer (2016) entry game example. Note when $\Pi^\star=\varnothing$ the value of $\boldsymbol{y}^\star$ does not matter, and so is marked with -.

(FC) Fixed strategic interaction parameters: $\delta_i^{LC}=\delta^{LC}$ and $\delta_i^{OA}=\delta^{OA}$.

(SR1) A sign restriction: $\delta_i^{LC},\delta_i^{OA}\leq 0$ (strategic substitutes).

(SR2) A sign restriction: $\beta_1^{LC},\beta_2^{LC},\beta_1^{OA},\beta_2^{OA}\geq 0$.

We then consider four counterfactual outcomes described in Table 5. The first counterfactual outcome $(KT1)$ is the average treatment effect on Y_i^{LC} of setting $Y_i^{OA}=1$ versus setting $Y_i^{OA}=0$. Similarly, the second counterfactual outcome $(KT2)$ is the average treatment effect on Y_i^{OA} of setting $Y_i^{LC}=1$ versus setting $Y_i^{LC}=0$. Both of these counterfactual outcomes measure the change in the distribution of the best response functions as the entry decision of the opposite player is changed.[45] The third counterfactual outcome $(KT3)$ is the average treatment effect on Y_i^{LC} of moving M_i^{OA} from its lowest to highest value, and the fourth counterfactual $(KT4)$ is the average treatment effect on Y_i^{OA} of moving M_i^{LC} from its lowest to highest value. Both $(KT3)$ and $(KT4)$ are relevant for evaluating the effectiveness of airline competition policy aimed at limiting the dominance of an airline (the proportion of routes serviced) in an airport.

The bounds on the four counterfactuals from Table 5 under various combinations of our assumptions are displayed in Table 6. Bounds under the weakest assumptions are displayed in the top line of Table 6, and bounds under the strongest assumptions are displayed in the bottom line of Table 6.

We find that the sign of counterfactuals $(KT1)$ and $(KT2)$ are identified and negative for all specifications that impose both independence restrictions and median zero restrictions, which is strong evidence in favor of a negative competition effect. Our narrowest bounds for $(KT1)$ are $[-0.68,-0.37]$, which is suggestive of a strong negative effect of the entry of the OA airlines on the entry of an LC airline. In contrast, the narrowest bounds for the outcome $(KT2)$ are $[-0.57,-0.02]$, which does not rule out small negative effects of the entry of an LC airlines on the entry of an OA airline. Unlike the bounds for $(KT1)$ and $(KT2)$ the bounds for $(KT3)$ and $(KT4)$ overlap zero under all combinations of assumptions, suggesting an ambiguous effect of

[45] Nonparametric identification of best response functions has also been studied by Kline and Tamer (2012) who provide analytical bounds under a fixed set of assumptions. Unlike our optimization-based approach, analytical bounds need to be re-derived each time the set of assumptions is changed.

Table 6: Results for the Kline and Tamer (2016) Airline Entry Game Example. The definition of each counterfactual outcome is provided in Table 5. The "Modal # of Cells" column indicates the modal (most frequent) number of latent space cells found by the cell enumeration algorithm, with the mode taken across all representative points.

(I)	*(EQ)*	*(MZ)*	*(FC)*	*(SR1)*	*(SR2)*	Representative Points	Modal # of Cells	*(KT1)*	*(KT2)*	*(KT3)*	*(KT4)*
						144	2401	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-1.00, 1.00]$	$[-1.00, 1.00]$
				✓		144	3136	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
				✓	✓	9	3136	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
			✓			57600	169	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
			✓	✓		14400	169	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
			✓	✓	✓	900	169	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
		✓				144	3136	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-1.00, 1.00]$	$[-1.00, 1.00]$
		✓		✓		144	4096	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
		✓		✓	✓	9	4096	$[-0.69, -0.27]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
		✓	✓			57600	196	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
		✓	✓	✓		14400	196	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
		✓	✓	✓	✓	900	196	$[-0.69, -0.27]$	$[-0.69, 0.00]$	$[-0.69, 0.69]$	$[-0.69, 0.69]$
	✓					144	2401	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.67, 0.67]$	$[-0.70, 0.64]$
	✓			✓		144	3136	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓			✓	✓	9	3136	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓		✓			57600	169	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓		✓	✓		14400	169	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓		✓	✓	✓	900	169	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓	✓				144	3136	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.67, 0.67]$	$[-0.7, 0.64]$
	✓	✓		✓		144	4096	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓	✓		✓	✓	9	4096	$[-0.69, -0.27]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓	✓	✓			57600	196	$[-0.69, 0.31]$	$[-0.69, 0.31]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓	✓	✓	✓		14400	196	$[-0.69, 0.00]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
	✓	✓	✓	✓	✓	900	196	$[-0.69, -0.27]$	$[-0.69, 0.00]$	$[-0.51, 0.41]$	$[-0.58, 0.36]$
✓						144	2401	$[-0.69, 0.29]$	$[-0.59, 0.18]$	$[-0.81, 0.82]$	$[-0.82, 0.85]$
✓				✓		144	3136	$[-0.69, 0.00]$	$[-0.59, -0.01]$	$[-0.63, 0.57]$	$[-0.59, 0.57]$
✓				✓	✓	9	3136	$[-0.69, 0.00]$	$[-0.59, -0.01]$	$[-0.63, 0.57]$	$[-0.59, 0.57]$
✓			✓			57600	169	$[-0.69, 0.23]$	$[-0.58, -0.01]$	$[-0.63, 0.51]$	$[-0.58, 0.50]$
✓			✓	✓		14400	169	$[-0.69, 0.00]$	$[-0.58, -0.01]$	$[-0.63, 0.51]$	$[-0.58, 0.50]$
✓			✓	✓	✓	900	169	$[-0.69, 0.00]$	$[-0.58, -0.01]$	$[-0.63, 0.51]$	$[-0.58, 0.50]$
✓		✓				144	3136	$[-0.69, 0.21]$	$[-0.59, 0.18]$	$[-0.81, 0.82]$	$[-0.82, 0.85]$
✓		✓		✓		144	4096	$[-0.69, -0.30]$	$[-0.59, -0.01]$	$[-0.63, 0.57]$	$[-0.59, 0.57]$
✓		✓		✓	✓	9	4096	$[-0.69, -0.30]$	$[-0.59, -0.01]$	$[-0.63, 0.57]$	$[-0.59, 0.57]$
✓		✓	✓			57600	196	$[-0.68, -0.37]$	$[-0.56, -0.04]$	$[-0.59, 0.47]$	$[-0.57, 0.50]$
✓		✓	✓	✓		14400	196	$[-0.68, -0.37]$	$[-0.58, -0.02]$	$[-0.63, 0.44]$	$[-0.57, 0.47]$
✓		✓	✓	✓	✓	900	196	$[-0.68, -0.37]$	$[-0.56, -0.04]$	$[-0.63, 0.49]$	$[-0.57, 0.39]$
✓	✓					144	2401	$[-0.69, 0.29]$	$[-0.59, 0.18]$	$[-0.57, 0.56]$	$[-0.62, 0.54]$
✓	✓			✓		144	3136	$[-0.69, 0.00]$	$[-0.59, -0.01]$	$[-0.48, 0.35]$	$[-0.48, 0.30]$
✓	✓			✓	✓	9	3136	$[-0.69, 0.00]$	$[-0.59, -0.01]$	$[-0.48, 0.35]$	$[-0.48, 0.30]$
✓	✓		✓			57600	169	$[-0.69, 0.23]$	$[-0.58, -0.01]$	$[-0.48, 0.31]$	$[-0.47, 0.28]$
✓	✓		✓	✓		14400	169	$[-0.69, 0.00]$	$[-0.58, -0.01]$	$[-0.48, 0.31]$	$[-0.47, 0.28]$
✓	✓		✓	✓	✓	900	169	$[-0.69, 0.00]$	$[-0.58, -0.01]$	$[-0.48, 0.31]$	$[-0.47, 0.28]$
✓	✓	✓				144	3136	$[-0.69, 0.21]$	$[-0.59, 0.18]$	$[-0.58, 0.56]$	$[-0.62, 0.54]$
✓	✓	✓		✓		144	4096	$[-0.69, -0.30]$	$[-0.59, -0.01]$	$[-0.48, 0.35]$	$[-0.48, 0.30]$
✓	✓	✓		✓	✓	9	4096	$[-0.69, -0.30]$	$[-0.59, -0.01]$	$[-0.48, 0.35]$	$[-0.48, 0.30]$
✓	✓	✓	✓			57600	196	$[-0.68, -0.37]$	$[-0.56, -0.03]$	$[-0.47, 0.28]$	$[-0.47, 0.22]$
✓	✓	✓	✓	✓		14400	196	$[-0.68, -0.37]$	$[-0.56, -0.03]$	$[-0.44, 0.27]$	$[-0.46, 0.22]$
✓	✓	✓	✓	✓	✓	900	196	$[-0.68, -0.37]$	$[-0.53, -0.03]$	$[-0.44, 0.27]$	$[-0.46, 0.22]$

any competition policy that seeks to limit airline dominance in a particular airport. The bounds for both $(KT3)$ and $(KT4)$ are $[-1.00, 1.00]$ in our baseline model, but narrow to $[-0.69, 0.69]$ when we impose our counterfactual equilibrium selection assumption, showing the assumption has strong identifying information. We find that the independence restriction and median zero restrictions are effective at reducing the width of the bounds for all counterfactual outcomes, especially when combined together. We also find a large range for the number of representative points, from 9 to $57,600$. The largest number of representative points occurs when both δ_i^{LC} and δ_i^{OA} are restricted to be constant across markets, and when there are no sign restrictions. This leads to a model with 6 fixed parameters with unrestricted signs. Our profiling procedure returns a geometric average of $(57,600)^{1/6} \approx 6.21$ points per dimension, equivalent to an extremely sparse grid although our procedure guarantees that relevant points of the parameter space have not been missed when computing counterfactual bounds. When $(SR1)$ is imposed the number of representative points drops from $57,600$ to $14,400$, indicating that $57,600 - 14,400 = 43,200$ (or exactly 75%) of the representative points have either $\delta^{LC} > 0$ or $\delta^{OA} > 0$. When $(SR2)$ is imposed the number of representative points drops further from $14,400$ to 900, indicating $14,400 - 900 = 13,400$ (or exactly 93.5%) of the representative points have either $\beta_1^{LC} < 0$, $\beta_2^{LC} < 0$, $\beta_1^{OA} < 0$ or $\beta_2^{OA} < 0$. Table 6 also illustrates that the modal number of cells (across representative points) increases if a sign restriction is imposed on a latent variable, since additional hyperplanes must be added to the initial arrangement to impose these restrictions.

6 Conclusion

In this paper we consider partial identification of a general class of counterfactual parameters in models with discrete outcome variables without parametric distributional assumptions on the latent variables. Our bounds are constructed by solving a sequence of linear programming problems, subject to a set of carefully constructed constraints. Our optimization formulation of the bounds allows the researcher to straightforwardly add and remove assumptions from the model. Implementation of our method requires a pre-processing step that first enumerates regions of the latent variable space, for which we propose a new cell enumeration algorithm. We also present a new algorithm for profiling structural parameters, which allows the researcher to construct exact (rather than approximate) bounds while using the sparsest grid possible over the fixed parameter space. We apply the method to a mobile phone choice example and an airline entry game example.

The methods in this paper add to a growing set of computational tools useful for the analysis of partially identified models, and we hope the theoretical results and discussion are useful for researchers studying counterfactuals in partially identified models. This paper continues a growing line of research that seeks to understand robustness of results to parametric distributional assumptions. Consistent with the conclusions of this literature, our applications show that parametric distributional assumptions play an important role in nonlinear models, especially when it comes to identifying counterfactual effects.

References

Aguiar, M. and Mahajan, S. (2017). *Topics in hyperplane arrangements*, volume 226. American Mathematical Soc.

Aguirregabiria, V. and Carro, J. (2021). Identification of average marginal effects in fixed effects dynamic discrete choice models.

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.

Andersen, E. D. (2009). The homogeneous and self-dual model and algorithm for linear optimization. Technical report, Technical Report TR-1-2009, MOSEK ApS, 2009.

Aradillas-López, A. (2020). The econometrics of static games. *Annual Review of Economics*, 12:135–165.

Aradillas-López, A. and Tamer, E. (2008). The identification power of equilibrium in simple games. *Journal of Business & Economic Statistics*, 26(3):261–283.

Aristodemou, E. (2021). Semiparametric identification in panel data discrete response models. *Journal of Econometrics*, 220(2):253–271.

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

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

Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. *Econometrica*, 78(5):1529–1568.

Balas, E. and Oosten, M. (1998). On the dimension of projected polyhedra. *Discrete Applied Mathematics*, 87(1-3):1–9.

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

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2021). Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048.

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

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica: Journal of the Econometric Society*, pages 1007–1028.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica: Journal of the Econometric Society*, pages 889–917.

Bjorn, P. A. and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation.

Borenstein, S. (1989). Hubs and high fares: dominance and market power in the us airline industry. *The RAND Journal of Economics*, pages 344–365.

Bresnahan, T. F. and Reiss, P. C. (1991a). Empirical models of discrete games. *Journal of Econometrics*, 48(1-2):57–81.

Bresnahan, T. F. and Reiss, P. C. (1991b). Entry and competition in concentrated markets. *Journal of Political Economy*, 99(5):977–1009.

Brueckner, J. K., Lee, D., and Singer, E. S. (2013). Airline competition and domestic us airfares: A comprehensive reappraisal. *Economics of Transportation*, 2(1):1–17.

Buck, R. C. (1943). Partition of space. *The American Mathematical Monthly*, 50(9):541–544.

Canen, N. and Song, K. (2020). A decomposition approach to counterfactual analysis in game-theoretic models. *arXiv preprint arXiv:2010.08868*.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

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. (2020). *Econometric Modeling of Interdependent Discrete Choice with Applications to Market Structure*. Cemmap, Centre for Microdata Methods and Practice, The Institute for Fiscal Studies.

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

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. (2022). Simple inference on functionals of set-identified parameters defined by linear moments. *arXiv preprint arXiv:1810.03180*.

Christensen, T. and Connault, B. (2019). Counterfactual sensitivity and robustness. *arXiv preprint arXiv:1904.00989*.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Ciliberto, F. and Williams, J. W. (2010). Limited access to airport facilities and market power in the airline industry. *The Journal of Law and Economics*, 53(3):467–495.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334.

Deb, R., Kitamura, Y., Quah, J. K.-H., and Stoye, J. (2018). Revealed price preference: theory and empirical analysis. *arXiv preprint arXiv:1801.02702*.

Dobronyi, C., Gu, J., and Kim, K. i. (2021). Identification of dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2104.04590*.

Dunker, F., Hoderlein, S., Kaido, H., and Sherman, R. (2018). Nonparametric identification of the distribution of random coefficients in binary response static games of complete information. *Journal of Econometrics*, 206(1):83–102.

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

Ekeland, I., Galichon, A., and Henry, M. (2010). Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory*, 42(2):355–374.

Evans, W. N. and Kessides, I. N. (1993). Localized market power in the us airline industry. *The Review of Economics and Statistics*, pages 66–75.

Fang, Z., Santos, A., Shaikh, A., and Torgovitsky, A. (2020). Inference for large-scale linear systems with known coefficients. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-134).

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *The RAND Journal of Economics*, 38(4):1002–1019.

Gafarov, B. (2019). Inference in high-dimensional set-identified affine models. *arXiv preprint arXiv:1904.00111*.

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

Goolsbee, A. and Syverson, C. (2008). How do incumbents respond to the threat of entry? evidence from the major airlines. *The Quarterly journal of economics*, 123(4):1611–1633.

Gu, J. and Koenker, R. (2020). Nonparametric maximum likelihood methods for binary response models with random coefficients. *Journal of the American Statistical Association*, pages 1–20.

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

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

Honoré, B. E. and de Paula, Á. (2021). Identification in simple binary outcome panel data models. *The Econometrics Journal*, 24(2):C78–C93.

Honoré, B. E. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74(3):611–629.

Jovanovic, B. (1989). Observable implications of models with multiple equilibria. *Econometrica: Journal of the Econometric Society*, pages 1431–1437.

Jun, S. J. and Pinkse, J. (2020). Counterfactual prediction in complete information games: point prediction under partial identification. *Journal of Econometrics*, 216(2):394–429.

Keane, M. and Wasi, N. (2013). Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6):1018–1045.

Khan, S., Ponomareva, M., and Tamer, E. (2021). Identification of dynamic binary response models. Technical report, Working paper.

Kitamura, Y. and Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909.

Kitamura, Y. and Stoye, J. (2019). Nonparametric counterfactuals in random utility models. *arXiv preprint arXiv:1902.08350*.

Kline, B. and Tamer, E. (2012). Bounds for best response functions in binary games. *Journal of Econometrics*, 166(1):92–105.

Kline, B. and Tamer, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics*, 7(2):329–366.

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

Lee, S. and Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, 86(6):1939–1963.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903.

Lu, Z. (2022). Estimating multinomial choice models with unobserved choice sets. *Journal of Econometrics*, 226(2):368–398.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228.

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

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

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

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

Molchanov, I. and Molinari, F. (2018). *Random sets in econometrics*, volume 60. Cambridge University Press.

Molinari, F. and Rosen, A. (2008). The identification power of equilibrium in games: The supermodular case. *Journal of Business & Economic Statistics*, 26(3):297–302.

Orlik, P. and Terao, H. (1992). *Arrangements of hyperplanes*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Germany, 1992 edition.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2015). Moment inequalities and their application. *Econometrica*, 83(1):315–334.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, pages 1029–1050.

Pearl, J. (2009). *Causality*. Cambridge University Press.

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

Schennach, S. M. (2014). Entropic latent variable integration via simulation. *Econometrica*, 82(1):345–385.

Sleumer, N. (1998). Output-sensitive cell enumeration in hyperplane arrangements. In Arnborg, S. and Ivansson, L., editors, *Algorithm Theory - SWAT'98*, pages 300–309. Springer-Verlag Berlin Heidelberg.

Stanley, R. P. (2004). An introduction to hyperplane arrangements. In *Lecture notes, IAS/Park City Mathematics Institute*.

Strotz, R. H. and Wold, H. O. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, pages 417–427.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165.

Tan, K. M. (2016). Incumbent response to entry by low-cost carriers in the us airline industry. *Southern Economic Journal*, 82(3):874–892.

Tebaldi, P., Torgovitsky, A., and Yang, H. (2021). Nonparametric estimates of demand in the california health insurance exchange. Technical report, National Bureau of Economic Research.

Torgovitsky, A. (2019a). Nonparametric inference on state dependence in unemployment. *Econometrica*, 87(5):1475–1505.

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

White, H. and Chalak, K. (2009). Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning. *Journal of Machine Learning Research*, 10(8).

Zaslavsky, T. (1975). *Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes*, volume 154. American Mathematical Soc.

Appendix A Proofs of Results in the Main Text

The Borel σ-algebra on a space $\mathcal{W}$ is denoted by $\mathfrak{B}(\mathcal{W})$. If $\mathcal{W} \subseteq \mathbb{R}^d$ contains a finite or countable number of elements, then the set is implicitly equipped with the discrete topology so that $\mathfrak{B}(\mathcal{W})$ corresponds to the power set of $\mathcal{W}$; otherwise, $\mathcal{W}$ is equipped with the usual topology. Given two $\sigma-$algebras $\mathfrak{A}_1$ and $\mathfrak{A}_2$ the product $\sigma-$algebra is denoted by $\mathfrak{A}_1 \otimes \mathfrak{A}_2$.

We are now prepared to prove the results in the main text.

Proof of Theorem 2.1. Suppose that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. Then by Definitions 2.1 and 2.2, there exists conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and some pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ satisfying:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{A.1}$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$, and:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{A.2}$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$. Using the conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and the pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta)$, define a conditional distribution $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$:

$$P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}_\ell} \int_{\mathcal{Y}} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \left(= \int_{\mathcal{U}_\ell} \int_{\mathcal{Y}} dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \right),$$

$P_{\boldsymbol{X}}-$almost surely for $\ell = 1, \ldots, L$.[46] Furthermore, define the joint distributions $P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}$ by setting:

$$P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}_\ell} \int_A dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{A.3}$$

and:

$$P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}_\ell} \int_A dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{A.4}$$

$P_{\boldsymbol{X}}-$almost surely, for $\ell = 1, \ldots, L$, and any $A \subset \mathcal{Y}$. By the Radon-Nikodym Theorem the existence of (a version of) both $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ is guaranteed since $P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}} \ll P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}} \ll P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability. Since all spaces involved are Euclidean, we can choose these versions to be almost surely unique regular conditional distributions (c.f. Durrett (2010) Theorem 5.1.9). Furthermore, by construction these satisfy:

$$P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{U}^\dagger \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_B \int_A dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{A.5}$$

and:

$$P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{U}^\dagger \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_B \int_A dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{A.6}$$

[46] The set $\mathcal{U}_\ell$ is a finite intersection of open half-spaces, and so is a Borel subset of $\mathcal{U}$.

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$ and $B \subset \mathcal{U}^\dagger$. Also note that:

$$\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) \text{ with } \boldsymbol{u} \in \mathcal{U}_\ell \iff \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \tag{A.7}$$

$$\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) \text{ with } \boldsymbol{u} \in \mathcal{U}_\ell \iff \boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta). \tag{A.8}$$

Finally, by Lemma B.1 the maps $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$ and $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$ are jointly measurable. Thus we have:

$$\begin{aligned}
&\int_{\mathcal{U}} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(1)}{=} \sum_{\ell=1}^{L} \int_{\mathcal{U}_\ell} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(2)}{=} \sum_{\ell=1}^{L} \int_{\mathcal{U}_\ell} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(3)}{=} \sum_{\ell=1}^{L} P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A \cap \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\qquad \overset{(4)}{=} \sum_{\ell=1}^{L} \int_{\{\boldsymbol{u}=\boldsymbol{u}_\ell\}} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\qquad \overset{(5)}{=} \int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}},
\end{aligned}$$

where the first equality follows from the fact that the distribution of $\boldsymbol{U}$ assigns zero probability to the hyperplanes by Assumption 2.1 and the definition of $\mathcal{U}_1, \ldots, \mathcal{U}_L$, the second equality follows from (A.7), the third equality follows from (A.3), the fourth equality follows from (A.5), and the fifth (and final) equality follows from the definition of $\mathcal{U}^\dagger$. Combining this with (A.1) shows that condition (2.13) holds $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$. In addition, we have:

$$\begin{aligned}
&\int_{\mathcal{U}} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(1)}{=} \sum_{\ell=1}^{L} \int_{\mathcal{U}_\ell} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(2)}{=} \sum_{\ell=1}^{L} \int_{\mathcal{U}_\ell} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\qquad \overset{(3)}{=} \sum_{\ell=1}^{L} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \cap \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\qquad \overset{(4)}{=} \sum_{\ell=1}^{L} \int_{\{\boldsymbol{u}=\boldsymbol{u}_\ell\}} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\qquad \overset{(5)}{=} \int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}},
\end{aligned}$$

where the first equality follows from the fact that the distribution of $\boldsymbol{U}$ assigns zero probability to the hyperplanes by Assumption 2.1 and the definition of $\mathcal{U}_1, \ldots, \mathcal{U}_L$, the second equality follows from (A.8), the

third equality follows from (A.4), the fourth equality follows from (A.6), and the fifth (and final) equality follows from the definition of $\mathcal{U}^\dagger$. Combining this with (A.2) shows that condition (2.14) holds $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$. This completes the proof of necessity.

For sufficiency, let $\boldsymbol{U}^\dagger$ be a random vector with support on $\mathcal{U}^\dagger$, and let $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$, and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ be any collections that satisfy conditions (2.13) and (2.14) $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$ for some $\theta \in \Theta$. We show that the joint distribution constructed from $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{X}}$, and the joint distribution constructed from $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{X}}$ can be extended to (not necessarily unique) Borel measures $P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{U}}$ on $\mathfrak{B}(\mathcal{Y})\otimes\mathfrak{B}(\mathcal{X})\otimes\mathfrak{B}(\mathcal{U})$ such that the associated conditional distribution $P_{\boldsymbol{U}|\boldsymbol{X}}$ satisfies $(P_{\boldsymbol{U}|\boldsymbol{X}},\theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ and such that (2.11) is satisfied. To this end, define the functions $\mu : \mathfrak{B}(\mathcal{Y})\otimes\mathfrak{B}(\mathcal{X})\otimes\mathfrak{B}(\mathcal{U}) \to \mathbb{R}$ and $\mu^\star : \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U}) \to \mathbb{R}$ as:

$$\begin{aligned}\mu(E) &:= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in E\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}}, \\ \mu^\star(E) &:= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y}^\star,\boldsymbol{x},\boldsymbol{u}) \in E\right\} dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}},\end{aligned} \tag{A.9}$$

where here $E \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$. We will now verify that μ is probability measure on $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$; the proof that $\mu^\star$ is a probability measure is nearly identical. To verify that μ is a probability measure, we must show that (i) $\mu(E) \geq \mu(\varnothing) = 0$ for every $E \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$, (ii) $\mu(\mathcal{Y} \times \mathcal{X} \times \mathcal{U}) = 1$, and (iii) for any countable sequence of disjoint sets $\{E_i\}_{i=1}^\infty$ in $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$, we have:

$$\mu\left(\bigcup_{i=1}^\infty E_i\right) = \sum_{i=1}^\infty \mu(E_i).$$

For property (i), first note that the quantity on the right side of (A.9) is always non-negative. Now note:

$$\mu(\varnothing) = \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in \varnothing\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} = 0.$$

Thus, property (i) is satisfied. For property (ii), note that:

$$\begin{aligned}\mu(\mathcal{Y} \times \mathcal{X} \times \mathcal{U}) &= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{U}\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} = 1.\end{aligned}$$

Finally, for property (iii), fix any countable disjoint sequence $\{E_i\}_{i=1}^\infty$ with $E_i \in \mathfrak{B}(\mathcal{Y})\otimes\mathfrak{B}(\mathcal{X})\otimes\mathfrak{B}(\mathcal{U})$. Then we have:

$$\begin{aligned}\mu\left(\bigcup_{i=1}^\infty E_i\right) &= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in \left(\bigcup_{i=1}^\infty E_i\right)\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \sum_{i=1}^\infty \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in E_i\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \\ &= \sum_{i=1}^\infty \int_{\mathcal{X}}\int_{\mathcal{U}^\dagger}\int_{\mathcal{Y}} \mathbb{1}\left\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in E_i\right\} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}}\end{aligned}$$

$$= \sum_{i=1}^{\infty} \mu\left(E_i\right),$$

where the second equality uses the fact that the sets $\{E_i\}_{i=1}^{\infty}$ are disjoint, and the third equality follows from Fubini's Theorem (by finiteness of μ). This verifies property (iii), and completes the proof that $\mu : \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U}) \to \mathbb{R}$ is a probability measure. An identical proof follows for $\mu^\star : \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U}) \to \mathbb{R}$.

Now set $P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}}$ equal to μ, $P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{U}}$ equal to $\mu^\star$. By the Radon-Nikodym Theorem, the existence of (a version of) $P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}}$ is guaranteed by the fact that $P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}} \ll P_{\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability, and the existence of (a version of) $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}$ is guaranteed by the fact that $P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{U}} \ll P_{\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability. Since all spaces involved are Euclidean, we can choose the versions $P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}$ to be almost surely unique regular conditional distributions (c.f. Durrett (2010) Theorem 5.1.9). Now define:

$$P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}) := \int_{\mathcal{Y}\times B} dP_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}} \left(= \int_{\mathcal{Y}\times B} dP_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}\right).$$

Again, by the Radon-Nikodym Theorem the existence of (a version of) $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ is guaranteed by the fact that $P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}} \ll P_{\boldsymbol{U}|\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability, and the existence of (a version of) $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ is guaranteed by the fact that $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}} \ll P_{\boldsymbol{U}|\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability. Since all spaces involved are Euclidean, we can choose the versions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ to be almost surely unique regular conditional distributions (c.f. Durrett (2010) Theorem 5.1.9). Furthermore, by construction $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ satisfy:

$$\begin{aligned} P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}) &= \int_B \int_A dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}, \\ P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}) &= \int_B \int_A dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}, \end{aligned} \tag{A.10}$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$ and $B \in \mathfrak{B}(\mathcal{U})$. We will now verify that $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$, and that (2.11) is satisfied. To do so, let us define the set:

$$\text{Graph}(\mathcal{Y}(\,\cdot\,,\theta)) := \{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) : \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)\}. \tag{A.11}$$

By Lemma B.1, this set is measurable. Now note that by construction we have for any $F \subset \mathcal{X}$:

$$\begin{aligned} &\int_{\mathcal{X}} P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}\right) \mathbb{1}\{\boldsymbol{x} \in F\}\, dP_{\boldsymbol{X}} \\ &\overset{(1)}{=} \int_{\mathcal{X}} \int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)\} \mathbb{1}\{\boldsymbol{x} \in F\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}}, \\ &\overset{(2)}{=} \int_{\mathcal{X}} \int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}) \in \text{Graph}(\mathcal{Y}(\,\cdot\,,\theta)) \cap A \times F \times \mathcal{U}^\dagger\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}}, \\ &\overset{(3)}{=} \mu\left(\text{Graph}(\mathcal{Y}(\,\cdot\,,\theta)) \cap A \times F \times \mathcal{U}^\dagger\right) \\ &\overset{(4)}{=} P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}}((\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}) \in \text{Graph}(\mathcal{Y}(\,\cdot\,,\theta)) \cap A \times F \times \mathcal{U}^\dagger) \end{aligned}$$

$$\overset{(5)}{=} \int_{\mathcal{X}} \int_{\mathcal{U}} \int_{\mathcal{Y}} \mathbb{1}\{(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) \in \text{Graph}(\mathcal{Y}(\,\cdot\,, \theta)) \cap A \times F \times \mathcal{U}^\dagger\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}\, dP_{\boldsymbol{X}},$$
$$\overset{(6)}{=} \int_{\mathcal{X}} \int_{\mathcal{U}^\dagger} \int_{A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \mathbb{1}\{\boldsymbol{x} \in F\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}\, dP_{\boldsymbol{X}}$$
$$\overset{(7)}{=} \int_{\mathcal{X}} \int_{\mathcal{U}} \int_{A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \mathbb{1}\{\boldsymbol{x} \in F\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}\, dP_{\boldsymbol{X}},$$

where the first equality follows from the fact that $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ satisfy condition (2.13) by assumption, the second equality follows from the definition of $\text{Graph}(\mathcal{Y}(\,\cdot\,, \theta))$ from (A.11), the third equality follows from the definition of μ, the fourth equality follows from the fact that we have set $P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}}$ equal to μ, the fifth equality follows from (A.10), the sixth equality also follows from the definition of $\text{Graph}(\mathcal{Y}(\,\cdot\,, \theta))$ from (A.11), and the seventh (and last) equality follows from the fact that $P_{\boldsymbol{U}|\boldsymbol{X}}$ concentrates on $\mathcal{U}^\dagger$ by construction. This last point also implies that the distribution $P_{\boldsymbol{U}} = P_{\boldsymbol{U}|\boldsymbol{X}} P_{\boldsymbol{X}}$ assigns zero probability to all hyperplanes from Assumption 2.1. Since the derivation above holds for any $F \subset \mathfrak{B}(\mathcal{X})$, conclude that:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_{A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}},$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$. By Definition 2.1, this implies $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$. Finally, following a similar proof as above, we can conclude that there exists a conditional distribution $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and a pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^*_{\boldsymbol{X}}$ satisfying:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_{A} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\}\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}},$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$. Applying Definition 2.2, this shows that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. This completes the proof of sufficiency, and thus completes the proof of Theorem 2.1.

■

Proof of Theorem 2.2. Throughout the proof we implicitly use Lemma B.1 to justify joint measurability of the functions $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$ and $(\boldsymbol{y}^\star, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$.

To prove validity (that is, that the proposed set from (2.18) contains the identified set), we must show that if there exists some collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ satisfying:

$$\overline{\Psi}_{\boldsymbol{x}} := \int_{\mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}},$$

almost surely, then $\overline{\Psi}_{\boldsymbol{x}}$ belongs to (2.18). Let us fix a collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ satisfying this condition. By Theorem 2.1, this implies that there exists a $\theta \in \Theta$, a collection of finite-dimensional conditional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ that satisfy conditions (2.13) and (2.14) $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$. We will show that, for this value of θ, we have $\overline{\Psi}_{\boldsymbol{x}} \in [\Psi^{\ell b}_{\boldsymbol{x}}(\theta), \Psi^{ub}_{\boldsymbol{x}}(\theta)]$ where the endpoints of this interval are determined by (2.19) and (2.20). To do so, note that it suffices to show that there exists a vector $\boldsymbol{\pi}$

satisfying (2.15), (2.16), and (2.17) (i.e. a feasible vector) such that:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}) = \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell). \tag{A.12}$$

This will be done by construction; in particular, let $\boldsymbol{\pi}$ be any probability vector satisfying the constraints:

$$\begin{aligned}
&\sum_{\boldsymbol{y}^\star\in\mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\qquad = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell) P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}), \qquad \text{(A.13)} \\
&\sum_{\boldsymbol{y}\in\mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\qquad = P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U}^\dagger = \boldsymbol{u}) P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}). \qquad \text{(A.14)}
\end{aligned}$$

Note that the set of vectors $\boldsymbol{\pi}$ satisfying these constraints is always nonempty.[47] Using (A.13) and (A.14), it is straightforward to verify $\boldsymbol{\pi}$ satisfies (2.16) and (2.17). Furthermore:

$$\begin{aligned}
&\sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star\in\mathcal{Y}} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(1)}{=} \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \sum_{\boldsymbol{y}^\star\in\mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(2)}{=} \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell) P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(3)}{=} \int_{\mathcal{U}^\dagger} \int_{\boldsymbol{y}} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(4)}{=} P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}\right),
\end{aligned}$$

where the second equality follows from (A.13), and the fourth equality follows from the fact that $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ satisfy (2.13). Conclude that $\boldsymbol{\pi}$ satisfies (2.15), and thus is feasible in the optimization problems (2.19) and (2.20). Finally, note that:

$$\begin{aligned}
&\sum_{\ell=1}^{L} \sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(1)}{=} \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \sum_{\boldsymbol{y}\in\mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(2)}{=} \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell) P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(3)}{=} \int_{\mathcal{U}^\dagger} \int_{\boldsymbol{y}^\star} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(4)}{=} P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}\right),
\end{aligned}$$

[47] For instance, we can always set $\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell)$ equal to the product of $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$, and $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$.

where the second equality follows from (A.14), and the fourth equality follows from the fact that $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ satisfy (2.14). Conclude that $\boldsymbol{\pi}$ satisfies (A.12). Thus we have shown that any probability vector $\boldsymbol{\pi}$ satisfying (A.13) and (A.14) is feasible in the optimization problems (2.19) and (2.20) and satisfies (A.12). By construction of the objective function in the optimization problems (2.19) and (2.20), this implies that $\overline{\Psi}_{\boldsymbol{x}} \in [\Psi^{\ell b}_{\boldsymbol{x}}(\theta), \Psi^{ub}_{\boldsymbol{x}}(\theta)]$ and completes the proof of validity.

We now show that, if $\overline{\Psi}_{\boldsymbol{x}}$ belongs to (2.18), then there exists some collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ satisfying:

$$\overline{\Psi}_{\boldsymbol{x}} := \int_{\mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}, \tag{A.15}$$

$P_{\boldsymbol{X}}-$almost surely. If $\overline{\Psi}_{\boldsymbol{x}}$ belongs to (2.18), then there exists $\theta \in \Theta$ such that $\overline{\Psi}_{\boldsymbol{x}} \in [\Psi^{\ell b}_{\boldsymbol{x}}(\theta), \Psi^{ub}_{\boldsymbol{x}}(\theta)]$, where the endpoints of this interval are determined by (2.19) and (2.20). Since the feasible region in the problems (2.19) and (2.20) are convex for this fixed θ, we know that there exists a vector $\boldsymbol{\pi}$ satisfying (2.15), (2.16), and (2.17) such that:

$$\overline{\Psi}_{\boldsymbol{x}} = \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell), \tag{A.16}$$

For this $\boldsymbol{\pi}$, consider setting:

$$P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) := \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell), \tag{A.17}$$

$$P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) := \sum_{\boldsymbol{y} \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell). \tag{A.18}$$

Furthermore, set:

$$P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) := \sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell),$$

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}) := \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell).$$

By the Radon-Nikodym Theorem the existence of (a version of) $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ is guaranteed by the fact that $P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}} \ll P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}} \ll P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ at all $\boldsymbol{X} = \boldsymbol{x}$ assigned positive probability. Since all spaces involved are Euclidean, we can choose these versions to be almost surely unique regular conditional distributions (c.f. Durrett (2010) Theorem 5.1.9). Furthermore, by construction $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ satisfy:

$$P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{U}^\dagger \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_B \int_A dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{A.19}$$

$$P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{U}^\dagger \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_B \int_A dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}}, \tag{A.20}$$

$P_{\boldsymbol{X}}-$almost surely for every $A \subset \mathcal{Y}$ and $B \subset \mathcal{U}^\dagger$. Using this, for any $A \subset \mathcal{Y}$ we have:

$$
\begin{aligned}
&\int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(1)}{=} \int_{\mathcal{U}^\dagger} \int_{A \cap \mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(2)}{=} \sum_{\ell=1}^{L} \int_{\{\boldsymbol{u}=\boldsymbol{u}_\ell\}} \int_{A \cap \mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)} dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(3)}{=} \sum_{\ell=1}^{L} P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A \cap \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(4)}{=} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(5)}{=} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(6)}{=} \sum_{\boldsymbol{y} \in A} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(7)}{=} \sum_{\boldsymbol{y} \in A} P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(8)}{=} P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}
$$

where the third equality follows from (A.19), the fifth equality follows from the definition of $P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}$ from (A.17), and the seventh equality follows from (2.15). Since this holds for any $\boldsymbol{x} \in \mathcal{X}$ assigned positive probability, this demonstrates that $P_{\boldsymbol{Y},\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ satisfy (2.13). Similarly, for any $A \subset \mathcal{Y}$ we have:

$$
\begin{aligned}
&\int_{\mathcal{U}^\dagger} \int_A \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(1)}{=} \int_{\mathcal{U}^\dagger} \int_{A \cap \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)} dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(2)}{=} \sum_{\ell=1}^{L} \int_{\{\boldsymbol{u}=\boldsymbol{u}_\ell\}} \int_{A \cap \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)} dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger} \, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}} \\
&\overset{(3)}{=} \sum_{\ell=1}^{L} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \cap \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta), \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(4)}{=} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star \in A} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{(5)}{=} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y}^\star \in A} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \sum_{\boldsymbol{y} \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(6)}{=} \sum_{\boldsymbol{y}^\star \in A} \sum_{\ell=1}^{L} \sum_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{u}_\ell) \\
&\overset{(7)}{=} \sum_{\boldsymbol{y}^\star \in A} P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x})
\end{aligned}
$$

$$\overset{(8)}{=} P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \mid \boldsymbol{X} = \boldsymbol{x}),$$

where the third equality follows from (A.20), the fifth equality follows from the definition of $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}$ from (A.18), and the seventh equality follows from the fact that $\boldsymbol{\pi}$ satisfies (2.17), which implies that:

$$\begin{aligned}\sum_{\ell=1}^{L}\sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u}_\ell,\theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell) &= \sum_{\ell=1}^{L}\sum_{\boldsymbol{y}\in\mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{y}^\star,\boldsymbol{u}_\ell)\\ &= \sum_{\ell=1}^{L} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star, \boldsymbol{U}^\dagger = \boldsymbol{u}_\ell \mid \boldsymbol{X} = \boldsymbol{x})\\ &= P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma = \boldsymbol{y}^\star \mid \boldsymbol{X} = \boldsymbol{x}).\end{aligned}$$

Since this holds for any $\boldsymbol{x} \in \mathcal{X}$ assigned positive probability, this demonstrates that $P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ satisfy (2.14) for the collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. Thus, we have shown by construction that there exists a $\theta \in \Theta$, a collection of conditional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$, and finite collections $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}^\dagger}$ that satisfy conditions (2.13) and (2.14) $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$ for the collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ above. Invoking Theorem 2.1, we conclude that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. Using (A.16), we see that this $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ also satisfies (A.15). Since $\overline{\Psi}_{\boldsymbol{x}}$ was arbitrary, we have shown that for any $\overline{\Psi}_{\boldsymbol{x}}$ belonging to (2.18), there exists a corresponding collection $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ satisfying (A.15). This completes the proof of Theorem 2.2.

■

Proof of Proposition 3.1. Fix any $\theta, \theta' \in \Theta$ satisfying $\mathcal{S}(\theta) = \mathcal{S}(\theta')$. By definition, this implies:

$$\{\boldsymbol{s} \in \{-1,+1\}^n : \mathcal{U}(\theta,\boldsymbol{s}) \neq \varnothing\} = \{\boldsymbol{s} \in \{-1,+1\}^n : \mathcal{U}(\theta',\boldsymbol{s}) \neq \varnothing\}.$$

By definition of $\mathcal{U}(\theta,\boldsymbol{s})$, this is equivalent to:

$$\left\{\left(\begin{array}{ll} & m_1 \in \mathcal{M}_1,\\ & m_2 \in \mathcal{M}_2,\\ & k = 1,\ldots,d_y,\\ \mathbb{1}\{g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\boldsymbol{\theta}) \geq 0\} : & A \in \mathcal{A}_k^{(m_1,m_2)},\\ & j \in A,\\ & \boldsymbol{y} \in \mathcal{Y},\\ & \boldsymbol{x} \in \mathcal{X}. \end{array}\right)\right\}_{\boldsymbol{u}\in\mathcal{U}.}$$

$$= \left\{ \left(\mathbb{1}\{g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\boldsymbol{\theta}') \geq 0\} : \begin{array}{l} m_1 \in \mathcal{M}_1, \\ m_2 \in \mathcal{M}_2, \\ k = 1,\ldots,d_y, \\ A \in \mathcal{A}_k^{(m_1,m_2)}, \\ j \in A, \\ \boldsymbol{y} \in \mathcal{Y}, \\ \boldsymbol{x} \in \mathcal{X}. \end{array} \right) \right\}_{\boldsymbol{u}\in\mathcal{U}} .$$

From the construction of $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)$ from Assumption 2.1 and 2.2, this implies:

$$\{(\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta),\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta))\}_{\boldsymbol{u}\in\mathcal{U}} = \{(\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta'),\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta'))\}_{\boldsymbol{u}\in\mathcal{U}},$$

for all $\boldsymbol{x} \in \mathcal{X}$. An identical proof shows that:

$$\{(\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta),\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta))\}_{\boldsymbol{u}\in\mathcal{U}} \subseteq \{(\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta'),\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta'))\}_{\boldsymbol{u}\in\mathcal{U}},$$

for all $\boldsymbol{x} \in \mathcal{X}$ when $\mathcal{S}(\theta) \subset \mathcal{S}(\theta')$.

■

Proof of Proposition 3.2. Without loss of generality suppose that there exists a cell $\mathcal{C}$ induced by $\mathscr{A}$ such that $\theta \in \text{Proj}_\Theta(\mathcal{C})$ and $\theta' \notin \text{Proj}_\Theta(\mathcal{C})$. Let $\boldsymbol{s}$ be the sign vector associated with this cell. Then $\theta \in \Theta(\boldsymbol{s})$ and $\theta' \notin \Theta(\boldsymbol{s})$, and by (3.5) we have $\mathcal{S}(\theta) \neq \mathcal{S}(\theta')$.

Now suppose $\mathcal{S}(\theta) \neq \mathcal{S}(\theta')$. Then either $\exists \boldsymbol{s} \in \mathcal{S}(\theta)$ such that $\boldsymbol{s} \notin \mathcal{S}(\theta')$ or $\exists \boldsymbol{s}' \in \mathcal{S}(\theta')$ such that $\boldsymbol{s}' \notin \mathcal{S}(\theta)$. Suppose it is the former. Then $\mathcal{U}(\theta,\boldsymbol{s}) \neq \varnothing$ and $\mathcal{U}(\theta',\boldsymbol{s}) = \varnothing$. Now let $\mathcal{C} = \mathcal{U}(\theta,\boldsymbol{s})$. Then $\theta \in \text{Proj}_\Theta(\mathcal{C})$ but $\theta' \notin \text{Proj}_\Theta(\mathcal{C})$. If instead $\exists \boldsymbol{s}' \in \mathcal{S}(\theta')$ such that $\boldsymbol{s}' \notin \mathcal{S}(\theta)$, then an identical argument shows there exists a cell $\mathcal{C}$ such that $\theta' \in \text{Proj}_\Theta(\mathcal{C})$ but $\theta \notin \text{Proj}_\Theta(\mathcal{C})$.

■

Proof of Proposition 3.3. Fix an arbitrary hyperplane $E \in \mathscr{E}(\mathcal{C})$, and let $\boldsymbol{h}_E^\top$ denote the normal vector for E. The vector $\boldsymbol{h}_E^\top$ contains at least one non-zero entry (otherwise E cannot define a half-space in the minimal representation of $\text{Proj}_\Theta(\mathcal{C})$). Without loss of generality, assume the first non-zero entry in $\boldsymbol{h}_E^\top$ is 1; if not, we can normalize $\boldsymbol{h}_E^\top$ so that this is true. Now consider the vector $\boldsymbol{h}^\top = (\boldsymbol{0}^\top, \boldsymbol{h}_E^\top)$ where $\boldsymbol{0}^\top$ is $1 \times d_u$. Furthermore, let H be the $d-1$ dimensional hyperplane with normal vector $\boldsymbol{h}^\top$. Clearly $E = \text{Proj}_\Theta(H)$. To show that H is a profiling hyperplane for $\mathcal{C}$, we verify the conditions of Definition 3.4. Condition (ii) is satisfied since the first d_u entries of $\boldsymbol{h}^\top$ are zero, and condition (iii) is satisfied since the first non-zero entry of $\boldsymbol{h}_E^\top$ is 1. It remains to verify conditions (i) and (iv). For condition (i), we must verify that $\mathcal{C}$ is contained in one of the open half-spaces defined by H, and that $\mathcal{C}$ has at least one boundary point on H. Recall that $\text{Proj}_\Theta(\mathcal{C})$ is contained in an open half-space defined by the hyperplane E. Without loss of generality, suppose

that it is the negative half-space; that is:

$$\text{Proj}_\Theta(\mathcal{C}) = \left\{\theta \in \mathbb{R}^{d_\theta} : \exists \boldsymbol{u} \in \mathbb{R}^{d_u} \text{ s.t. } (\boldsymbol{u}, \theta) \in \mathcal{C}\right\} \subset \{\theta \in \mathbb{R}^{d_\theta} : \langle \boldsymbol{h}_E, \theta\rangle < 0\}. \tag{A.21}$$

Now fix an arbitrary $\tilde{\boldsymbol{v}} = (\tilde{\boldsymbol{u}}, \tilde{\theta}) \in \mathcal{C}$. Then clearly $\tilde{\theta} \in \text{Proj}_\Theta(\mathcal{C})$, and thus:

$$\langle \boldsymbol{h}, \tilde{\boldsymbol{v}}\rangle = \langle \boldsymbol{h}_E, \tilde{\theta}\rangle < 0,$$

where the first equality from the definition of $\boldsymbol{h}$, and the last inequality follows from the fact that $\tilde{\theta} \in \text{Proj}_\Theta(\mathcal{C})$ and (A.21). Since $\tilde{\boldsymbol{v}} \in \mathcal{C}$ was arbitrary, conclude that:

$$\mathcal{C} \subset \left\{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{h}, \boldsymbol{v}\rangle < 0\right\}. \tag{A.22}$$

Now note that, by construction, the hyperplane E is a supporting hyperplane of $\text{Proj}_\Theta(\mathcal{C})$. Thus, there exists at least one $\check{\theta} \in \partial\text{Proj}_\Theta(\mathcal{C})$ such that $\langle \boldsymbol{h}_E, \check{\theta}\rangle = 0$. Let $\{\check{\theta}_n\}_{n=1}^\infty \subset \text{Proj}_\Theta(\mathcal{C})$ be a sequence in $\text{Proj}_\Theta(\mathcal{C})$ converging to $\check{\theta}$. Since $\{\check{\theta}_n\}_{n=1}^\infty \subset \text{Proj}_\Theta(\mathcal{C})$, by definition of the projection we have for each $\check{\theta}_n$ there exists a corresponding $\check{\boldsymbol{u}}_n$ such that $\check{\boldsymbol{v}}_n := (\check{\boldsymbol{u}}_n, \check{\theta}_n) \in \mathcal{C}$. Since $\mathcal{C}$ is a cone, we can assume without loss of generality that $||\check{\boldsymbol{v}}_n|| \leq 1$; otherwise, redefine $\check{\boldsymbol{v}}_n \leftarrow \check{\boldsymbol{v}}_n / \max\{1, ||\check{\boldsymbol{v}}_n||\}$. Since $||\check{\boldsymbol{v}}_n|| \leq 1$, the sequence $\{\check{\boldsymbol{v}}_n\}_{n=1}^\infty$ is bounded, and so admits a convergent subsequence, say $\{\check{\boldsymbol{v}}_{n_k}\}_{k=1}^\infty \to \check{\boldsymbol{v}}$. Also, since $\{\check{\boldsymbol{v}}_n\}_{n=1}^\infty \subset \mathcal{C}$, we must have $\check{\boldsymbol{v}} \in \text{cl}(\mathcal{C})$. Now note that the distance between $\check{\boldsymbol{v}}_{n_k}$ and the hyperplane H is given by:

$$\frac{|\langle \boldsymbol{h}, \check{\boldsymbol{v}}_{n_k}\rangle|}{||\boldsymbol{h}||} \to \frac{|\langle \boldsymbol{h}, \check{\boldsymbol{v}}\rangle|}{||\boldsymbol{h}||} = \frac{|\langle \boldsymbol{h}_E, \check{\theta}\rangle|}{||\boldsymbol{h}||} = 0, \tag{A.23}$$

which follows since $\{\check{\theta}_{n_k}\}_{k=1}^\infty \subset \{\check{\theta}_n\}_{n=1}^\infty \to \check{\theta}$. This shows that $\check{\boldsymbol{v}} \in H$. Furthermore, we know that $\check{\boldsymbol{v}} \in \text{cl}(\mathcal{C})$, but if $\check{\boldsymbol{v}} \in \mathcal{C}$ then (A.23) would contradict (A.22). Conclude that $\check{\boldsymbol{v}} \in \partial\mathcal{C}$ and $\check{\boldsymbol{v}} \in H$. This result shows $\mathcal{C}$ has at least one boundary point on the hyperplane H, and (A.22) shows that $\mathcal{C}$ is contained in the open half-space defined by H. Together, this shows that H is a supporting hyperplane of $\mathcal{C}$ and verifies condition (i).

Finally, we will verify H satisfies condition (iv) in Definition 3.4. Since H is a supporting hyperplane of $\mathcal{C}$, it must contain the face:

$$F = \{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \langle \boldsymbol{h}, \boldsymbol{v}\rangle = 0\} = \{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \tilde{\boldsymbol{A}}_u \boldsymbol{u} + \tilde{\boldsymbol{A}}_\theta \theta = 0\},$$

where $\tilde{\boldsymbol{A}}_u = \boldsymbol{0}^\top$ and $\tilde{\boldsymbol{A}}_\theta = \boldsymbol{h}_E^\top$, so that $\text{rank}(\tilde{\boldsymbol{A}}_u) = 0$ and $\text{rank}(\tilde{\boldsymbol{A}}_\theta) = 1$ (since $\boldsymbol{h}_E^\top$ contains at least one non-zero element, as argued earlier). Thus, H contains a face of $\mathcal{C}$ defined by the matrix $\tilde{\boldsymbol{A}} = [\tilde{\boldsymbol{A}}_u, \tilde{\boldsymbol{A}}_\theta]$ satisfying $1 \leq \text{rank}(\tilde{\boldsymbol{A}}) \leq d_u + 1$ and $\text{rank}(\tilde{\boldsymbol{A}}) - \text{rank}(\tilde{\boldsymbol{A}}_u) = 1$, which verifies condition (iv) in Definition 3.4. Combining everything, conclude that H is a profiling hyperplane.

For the reverse direction, we show that every profiling hyperplane for $\mathcal{C}$ contains one of the hyperplanes in $\mathscr{E}(\mathcal{C})$. Let H be a profiling hyperplane for $\mathcal{C}$. By conditions (ii) and (iii) the normal vector $\boldsymbol{a}$ for H must contain zeros in the first d_u entries, and a leading 1 in the last $d - d_u$ entries. By condition (i), H is a

supporting hyperplane for $\mathcal{C}$, and so must contain the face:

$$F := \{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0\}.$$

Since the first d_u entries of $\boldsymbol{a}$ are zero, by Proposition 3.4 in Balas and Oosten (1998) $\text{Proj}_\Theta(F)$ is a face of $\text{Proj}_\Theta(\mathcal{C})$. Since H contains F, $\text{Proj}_\Theta(H)$ contains $\text{span}(\text{Proj}_\Theta(F)))$. Now write $\boldsymbol{a}^\top = (\boldsymbol{a}_u^\top, \boldsymbol{a}_\theta^\top)$ where $\boldsymbol{a}_u^\top$ is $1 \times d_u$ and $\boldsymbol{a}_\theta^\top$ is $1 \times d_\theta$. By Theorem 2.4 in Balas and Oosten (1998):

$$\dim(\text{Proj}_\Theta(F)) = \dim(F) - d_u + \text{rank}(\boldsymbol{a}_u) = d_\theta - \text{rank}(\boldsymbol{a}_\theta) + \text{rank}(\boldsymbol{a}_u) = d_\theta - 1 + 0 = d_\theta - 1.$$

Since every $d_\theta - 1$ dimensional face of $\text{Proj}_\Theta(\mathcal{C})$ is contained in exactly one of the hyperplanes in $\mathscr{E}(\mathcal{C})$, the face $\text{Proj}_\Theta(F)$ is contained in exactly one of these hyperplanes, say E. Since the dimension of $\text{Proj}_\Theta(F)$ is $d_\theta - 1$, it must be that $E = \text{span}(\text{Proj}_\Theta(F))$. But since $\text{Proj}_\Theta(H)$ contains $\text{span}(\text{Proj}_\Theta(F))$, this shows $\text{Proj}_\Theta(H)$ contains one of the hyperplanes in $\mathscr{E}(\mathcal{C})$. Since H was an arbitrary profiling hyperplane, conclude that each profiling hyperplane for $\mathcal{C}$ contains one of the hyperplanes in $\mathscr{E}(\mathcal{C})$. ■

Proof of Proposition 3.4. First we show that, for a given cell $\mathcal{C}$ induced by the arrangement $\mathscr{A}$, the rows of $\boldsymbol{A}(\mathcal{C})$ (the normals of all profiling hyperplanes of $\mathcal{C}$) are contained in $\boldsymbol{V}$. To this end, let $\boldsymbol{A}$ be the $n \times d$ matrix whose rows are normal vectors of the hyperplanes in the arrangement $\mathscr{A}$. Fix a cell $\mathcal{C}$ induced by the arrangement $\mathscr{A}$, and suppose that $\boldsymbol{a}$ is the normal vector of a profiling hyperplane of $\mathcal{C}$. Since this hyperplane is orthogonal to $\mathbb{R}^{d_\theta}$, its normal vector is orthogonal to $\mathbb{R}^{d_u}$, which can hold only if the first d_u elements of $\boldsymbol{a}$ are zero. Furthermore, by condition (iv) in Definition 3.4, there exists some $1 \leq j \leq d_u + 1$ and some matrix $\tilde{\boldsymbol{A}}$ with $\text{rank}(\tilde{\boldsymbol{A}}) = j$ (whose rows are contained in $\boldsymbol{A}$) and $\text{rank}(\tilde{\boldsymbol{A}}) - \text{rank}(\tilde{\boldsymbol{A}}_u) = 1$ such that $\{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\} \subseteq \{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{v} \rangle = 0\}$.

We now show that this implies $\{\boldsymbol{v} \in \mathbb{R}^d : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\} \subseteq \{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{v} \rangle = 0\}$. To do so, let $\tilde{n} = \text{nrow}(\tilde{\boldsymbol{A}})$, and note that:

$$\{\boldsymbol{v} \in \mathbb{R}^d : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\} = \text{span}\{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\}.$$

Now take any $\boldsymbol{v}' \in \{\boldsymbol{v} \in \mathbb{R}^d : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\}$. By the previous display, there exists a finite collection of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K \in \{\boldsymbol{v} \in \text{cl}(\mathcal{C}) : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\}$ and scalars $t_1, \ldots, t_K$ such that:

$$\boldsymbol{v}' = \sum_{k=1}^{K} t_k \boldsymbol{v}_k.$$

But then clearly:

$$\langle \boldsymbol{a}, \boldsymbol{v}' \rangle = \left\langle \boldsymbol{a}, \left(\sum_{k=1}^{K} t_k \boldsymbol{v}_k \right) \right\rangle = \sum_{k=1}^{K} t_k \langle \boldsymbol{a}, \boldsymbol{v}_k \rangle = 0.$$

Since $\boldsymbol{v}'$ was arbitrary we can conclude that $\{\boldsymbol{v} \in \mathbb{R}^d : \tilde{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}\} \subseteq \{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{v} \rangle = 0\}$, as claimed. This shows that the null space of $\tilde{\boldsymbol{A}}$ is contained within the null space of the row vector $\boldsymbol{a}^\top$. By the fundamental

theorem of linear algebra, this is true if and only if the column space of $\tilde{\boldsymbol{A}}^\top$ contains the vector $\boldsymbol{a}$. In other words, $\boldsymbol{a}^\top$ must be a linear combination of the rows of $\tilde{\boldsymbol{A}}$.

We now claim that $\boldsymbol{a}^\top$ must be a row in the reduced row echelon form of $\tilde{\boldsymbol{A}}$. Without loss of generality, assume $\tilde{\boldsymbol{A}}$ is in reduced row echelon form. Since $\boldsymbol{a}^\top$ lies in the span of the rows of $\tilde{\boldsymbol{A}}$, say (in order) $\tilde{\boldsymbol{a}}_1^\top, \tilde{\boldsymbol{a}}_2^\top, \ldots, \tilde{\boldsymbol{a}}_{\tilde{n}}^\top$, there exist scalars $t_1, \ldots, t_{\tilde{n}}$ such that:

$$\boldsymbol{a}^\top = t_1\tilde{\boldsymbol{a}}_1^\top + t_2\tilde{\boldsymbol{a}}_2^\top + \ldots + t_{\tilde{n}}\tilde{\boldsymbol{a}}_{\tilde{n}}^\top. \tag{A.24}$$

Since $\tilde{\boldsymbol{A}}$ is in reduced row echelon form, then for $k = 1, \ldots, \tilde{n}$, either the leading entry of $\tilde{\boldsymbol{a}}_k^\top$ is 1, or $\tilde{\boldsymbol{a}}_k^\top$ is the zero vector. Let $\tilde{k}$ denote the number of vectors with a leading 1 in one of the first d_u positions. Since the leading entry of $\boldsymbol{a}^\top$ is 1, and since $\boldsymbol{a}^\top$ contains d_u zeroes as its first entries, it must be that $t_k = 0$ for all $k \leq \tilde{k}$. Consider the following partition of $\tilde{\boldsymbol{A}}$:

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}}_{11} & \tilde{\boldsymbol{A}}_{12} \\ \tilde{\boldsymbol{A}}_{21} & \tilde{\boldsymbol{A}}_{22} \end{bmatrix},$$

where $\tilde{\boldsymbol{A}}_{11}$ is $\tilde{k} \times d_u$, $\tilde{\boldsymbol{A}}_{21}$ is $(\tilde{n} - \tilde{k}) \times d_u$, $\tilde{\boldsymbol{A}}_{12}$ is $\tilde{k} \times d_\theta$, and $\tilde{\boldsymbol{A}}_{22}$ is $(\tilde{n} - \tilde{k}) \times d_\theta$. Since $\tilde{\boldsymbol{A}}$ is in reduced row echelon form, and the first $\tilde{k}$ rows of $\tilde{\boldsymbol{A}}$ contain a leading 1 in the first d_u positions, it must be that $\tilde{\boldsymbol{A}}_{21} = \boldsymbol{0}$, and it must be that $\text{rank}(\tilde{\boldsymbol{A}}_u) = \text{rank}(\tilde{\boldsymbol{A}}_{11}) = \text{rank}([\tilde{\boldsymbol{A}}_{11}\ \tilde{\boldsymbol{A}}_{12}]) = \tilde{k}$ (since $\tilde{k} \leq d_u$). Now recall:

$$\tilde{\boldsymbol{A}}_u = \begin{bmatrix} \tilde{\boldsymbol{A}}_{11} \\ \tilde{\boldsymbol{A}}_{21} \end{bmatrix}.$$

Since $\text{rank}(\tilde{\boldsymbol{A}}) - \text{rank}(\tilde{\boldsymbol{A}}_u) = 1$, it must be that $\text{rank}([\tilde{\boldsymbol{A}}_{21}\ \tilde{\boldsymbol{A}}_{22}]) = \text{rank}(\tilde{\boldsymbol{A}}_{22}) = 1$. But since $\tilde{\boldsymbol{A}}$ is in reduced row echelon form, this is possible only if the last $\tilde{n} - \tilde{k} - 1$ rows of $\tilde{\boldsymbol{A}}$ are zero. Thus, for $k > \tilde{k} + 1$, we can take $t_k = 0$ in (A.24). But then we have shown that $t_k = 0$ for $1 \leq k \leq \tilde{k}$ and for $k > \tilde{k} + 1$. Conclude from (A.24) (and the fact that the leading entry of $\boldsymbol{a}^\top$ is 1) that $\boldsymbol{a}^\top$ must be equal to the $(\tilde{k} + 1)^{th}$ row of $\tilde{\boldsymbol{A}}$.

For the reverse, let $\boldsymbol{a}^\top$ be any row belonging to the matrix $\boldsymbol{V}$ returned by Algorithm 3. We will demonstrate that there exists a cell $\mathcal{C}$ induced by the arrangement $\mathscr{A}$ such that $\boldsymbol{a}$ is the normal vector of a profiling hyperplane for $\mathcal{C}$. Note that Algorithm 3 will not include $\boldsymbol{a}^\top$ as a row in $\boldsymbol{V}$ unless its first d_u entries are zero, and its leading entry is 1, so we know $\boldsymbol{a}^\top$ satisfies conditions (ii) and (iii) of Definition 3.4.

To verify that $\boldsymbol{a}^\top$ satisfies condition (i) in Definition 3.4, we must show that it is the supporting hyperplane for some cell $\mathcal{C}$ induced by the arrangement. If $\boldsymbol{a}^\top$ belongs to $\boldsymbol{V}$, then it is the linear combination of the rows of some matrix $\tilde{\boldsymbol{A}}$ whose rows are contained in $\boldsymbol{A}$. Suppose that $\tilde{\boldsymbol{A}}$ has $r \leq d$ rows (Algorithm 3 ensures this is the case), and rank $\tilde{r} \leq r$. Consider two cases:

1. $\tilde{r} < d$: Drop any rows of $\tilde{\boldsymbol{A}}$ that are linearly dependent, and suppose $r_1 \leq r$ rows remain. Of the remaining rows, suppose that $\boldsymbol{a}^\top$ is a linear combination of the first $r_2 \leq r_1$ rows $\tilde{\boldsymbol{a}}_1^\top, \ldots, \tilde{\boldsymbol{a}}_{r_2}^\top$ (otherwise re-label the rows to ensure this is the case). Let the vectors $\tilde{\boldsymbol{a}}_1^\top, \ldots, \tilde{\boldsymbol{a}}_{r_2}^\top$ represent the normal vectors of a hyperplane arrangement in $\mathbb{R}^d$. By construction, $r_2 < d$ and these vectors are linearly

independent, so the resulting arrangement is in general position. Adding $\boldsymbol{a}$ to this arrangement, the resulting arrangement cannot be in general position, since $\boldsymbol{a}^\top$ is a linear combination of $\tilde{\boldsymbol{a}}_1^\top, \ldots, \tilde{\boldsymbol{a}}_{r_2}^\top$. In particular, the hyperplane with normal vector $\boldsymbol{a}$ contains the intersection of the hyperplanes defined by the normal vectors $\tilde{\boldsymbol{a}}_1, \ldots, \tilde{\boldsymbol{a}}_{r_2}$, which is a $d - r_2 \geq 1$ dimensional subspace. This implies that the hyperplane with normal vector $\boldsymbol{a}$ cannot intersect all cells defined by the arrangement $\tilde{\boldsymbol{a}}_1, \ldots, \tilde{\boldsymbol{a}}_{r_2}$. Thus, at least one cell lies in a closed half-space defined by $\boldsymbol{a}$. Since this is a linear arrangement, all cells contain the origin as a boundary point. Since $\boldsymbol{a}$ passes through the origin, this verifies that $\boldsymbol{a}$ contains a boundary point of all cells. Since at least one cell lies in a closed half-space defined by $\boldsymbol{a}$, it must be that $\boldsymbol{a}$ is the normal vector of a supporting hyperplane for some cell.

2. $\tilde{r} = d$: Since $\tilde{r} \leq r \leq d$, this implies $\tilde{r} = r = d$. In this case, $\tilde{\boldsymbol{A}}$ has full rank, and the rows of $\tilde{\boldsymbol{A}}$ are normal vectors of a hyperplane arrangement in $\mathbb{R}^d$ in general position. Using the counting formula in (3.2), there are:

$$2\sum_{k=0}^{d-1} \binom{r-1}{k} = 2^d = 2^r = 2^{\tilde{r}},$$

cells induced by this arrangement. Adding the hyperplane defined by the normal vector $\boldsymbol{a}^\top$ to this arrangement, we have by the same formula:

$$2\sum_{k=0}^{d-1} \binom{(r+1)-1}{k} < 2^{r+1}.$$

But this shows that the hyperplane defined by $\boldsymbol{a}^\top$ cannot cross all cells in the arrangement defined by the rows of $\tilde{\boldsymbol{A}}$, since otherwise we would have 2^{r+1} cells. Following a similar argument to the first case, conclude that $\boldsymbol{a}^\top$ is the normal vector of a supporting hyperplane for some cell.

It remains to show that $\boldsymbol{a}^\top$ satisfies condition (iv). By the previous part, $\boldsymbol{a}^\top$ is the normal vector of a supporting hyperplane for some cell $\mathcal{C}$. This implies that the hyperplane defined by $\boldsymbol{a}^\top$ contains the face:

$$\{\boldsymbol{v} \in \mathrm{cl}(\mathcal{C}) : \langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0\}.$$

Since $\boldsymbol{a}^\top$ is a row in the reduced row echelon form of some matrix $\tilde{\boldsymbol{A}}$ (whose rows are contained in $\boldsymbol{A}$), we have by the fundamental theorem of linear algebra:

$$\{\boldsymbol{v} \in \mathrm{cl}(\mathcal{C}) : \tilde{\boldsymbol{A}}\boldsymbol{v} = 0\} \subseteq \{\boldsymbol{v} \in \mathrm{cl}(\mathcal{C}) : \langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0\} \subseteq \{\boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{v}, \boldsymbol{a} \rangle = 0\}. \tag{A.25}$$

Note that the set on the left side of the previous display is nonempty (it contains $\boldsymbol{v} = \boldsymbol{0}$), and so is a face of $\mathcal{C}$. It remains to show that $\mathrm{rank}(\tilde{\boldsymbol{A}}) - \mathrm{rank}(\tilde{\boldsymbol{A}}_u) = 1$.

As an intermediate step we now show that, if the reduced row echelon form of a submatrix $\boldsymbol{B}$ of $\boldsymbol{A}$ leads to a row $\boldsymbol{a}^\top$ that will be included in $\boldsymbol{V}$, then Algorithm 3 will ensure that $\boldsymbol{a}^\top$ is the only row in the reduced row echelon form of $\boldsymbol{B}$ that has a leading 1 outside the first d_u positions. Suppose not; that is, suppose

that Algorithm 3 reaches a matrix $\boldsymbol{B}$ such that the reduced row echelon form of $\boldsymbol{B}$ has multiple rows with a leading 1 outside the first d_u positions. Without loss of generality suppose the non-zero rows $\boldsymbol{a}_1^\top$ and $\boldsymbol{a}_2^\top$ from the reduced row echelon form of the matrix $\boldsymbol{B}$ both have a leading 1 outside the first d_u positions, and that the leading 1 in $\boldsymbol{a}_1^\top$ is in a position before that of the leading 1 in $\boldsymbol{a}_2^\top$. Both $\boldsymbol{a}_1^\top$ and $\boldsymbol{a}_2^\top$ lie in the linear span of the rows of $\boldsymbol{B}$, say (in order) $\boldsymbol{b}_1^\top, \boldsymbol{b}_2^\top, \ldots, \boldsymbol{b}_K^\top$, so that there exists coefficients $\{\lambda_k\}_{k=1}^K$ and $\{\tau_k\}_{k=1}^K$ (not all equal to zero) such that:

$$\begin{aligned}\boldsymbol{a}_1^\top &= \lambda_1 \boldsymbol{b}_1^\top + \lambda_2 \boldsymbol{b}_2^\top + \cdots + \lambda_K \boldsymbol{b}_K^\top, \\ \boldsymbol{a}_2^\top &= \tau_1 \boldsymbol{b}_1^\top + \tau_2 \boldsymbol{b}_2^\top + \cdots + \tau_K \boldsymbol{b}_K^\top.\end{aligned}$$

Suppose without loss of generality that $\tau_K \neq 0$. Then we have:

$$\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top = \left(\lambda_1 - \frac{\tau_1\lambda_K}{\tau_K}\right)\boldsymbol{b}_1^\top + \left(\lambda_2 - \frac{\tau_2\lambda_K}{\tau_K}\right)\boldsymbol{b}_2^\top + \cdots + \left(\lambda_{K-1} - \frac{\tau_{K-1}\lambda_K}{\tau_K}\right)\boldsymbol{b}_{K-1}^\top.$$

Since both $\boldsymbol{a}_1^\top$ and $\boldsymbol{a}_2^\top$ have their first d_u entries as zero, $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top$ will also have their first d_u entries as zero. Furthermore, $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top \neq \boldsymbol{0}$: if $\lambda_K = 0$ then $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top = \boldsymbol{0}$ would imply $\boldsymbol{a}_1^\top = \boldsymbol{0}$, a contradiction to the fact $\boldsymbol{a}_1^\top$ is non-zero, and if $\lambda_K \neq 0$ then $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top = \boldsymbol{0}$ implies that $\boldsymbol{a}_1^\top$ is a scalar multiple of $\boldsymbol{a}_2^\top$, contradicting the fact that $\boldsymbol{a}_1^\top$ and $\boldsymbol{a}_2^\top$ are non-zero rows from the reduced row echelon form of $\boldsymbol{B}$. Conclude that the vector $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top$ lies in the linear span of a matrix $\tilde{\boldsymbol{B}}$ formed by the first $K-1$ rows of the matrix $\boldsymbol{B}$. But this means the reduced row echelon form of $\tilde{\boldsymbol{B}}$ must contain a non-zero row—the vector $\boldsymbol{a}_1^\top - \frac{\lambda_K}{\tau_K}\boldsymbol{a}_2^\top$—that has a leading 1 in none of the first d_u positions. Such a row will be admitted to $\boldsymbol{V}$ at a previous iteration, and step (3b) Algorithm 3 ensures that we will never do row reduction on the matrix $\boldsymbol{B}$. Since the matrix $\boldsymbol{B}$ can never be reached by Algorithm 3, this contradicts our initial statement that $\boldsymbol{B}$ has multiple rows with a leading 1 outside the first d_u positions, and completes the proof of our intermediate step.

With this established, we can finish the proof. Recall that the vector $\boldsymbol{a}^\top$ is included in a row of $\boldsymbol{V}$ if and only if it is a non-zero row in the reduced row echelon form of some matrix $\tilde{\boldsymbol{A}}$ (whose rows are taken from $\boldsymbol{A}$), and has a leading 1 in none of the first d_u positions. Algorithm 3 also ensures that $\text{nrow}(\tilde{\boldsymbol{A}}) \leq d_u + 1$, which implies that $\text{rank}(\tilde{\boldsymbol{A}}) \leq d_u + 1$ (since $\tilde{\boldsymbol{A}}$ has $d = d_u + d_\theta$ columns, and Algorithm 3 only accepts matrices with $d_u, d_\theta \geq 1$). Since $\boldsymbol{a}^\top$ is the only row that has a leading 1 in none of the first d_u positions (by the intermediate step above), and the submatrix $\tilde{\boldsymbol{A}}_u$ contains just the first d_u columns of $\tilde{\boldsymbol{A}}$, we have $\text{rank}(\tilde{\boldsymbol{A}}) - \text{rank}(\tilde{\boldsymbol{A}}_u) = 1$. Combining this with (A.25), we can conclude that $\boldsymbol{a}$ satisfies condition (iv) in Definition 3.4. This verifies that $\boldsymbol{a}$ is the normal vector of a profiling hyperplane, and completes the proof.

■

Proof of Proposition 3.5. Suppose Algorithm 3 reaches step 3 with $j = d_u + 1$ without discarding any rows of $\boldsymbol{A}$ up to that point (we will prove this is the case below). At this step Algorithm 3 will consider all $\binom{n}{d_u+1}$ possible ways of intersecting of $d_u + 1$ hyperplanes from the arrangement $\mathscr{A}$. Because all hyperplanes are

in general position in $\mathbb{R}^d$, the intersection of each of these $d_u + 1$ hyperplanes form a $d - (d_u + 1) = d_\theta - 1$ hyperplane. We now show that this intersection is contained in $\boldsymbol{V}$.

Let the matrix $\boldsymbol{B}$ be of dimension $(d_u + 1) \times d$ whose rows (in order) are $\boldsymbol{b}_1^\top, \boldsymbol{b}_2^\top, \ldots, \boldsymbol{b}_{d_u+1}^\top$ and are taken from the rows of $\boldsymbol{A}$ (the matrix input to Algorithm 3). Recall that $\boldsymbol{v}^\top = (\boldsymbol{u}^\top, \theta^\top)$, and consider the system of equations:

$$\begin{aligned} \boldsymbol{b}_1^\top \boldsymbol{v} &= 0, \\ \boldsymbol{b}_2^\top \boldsymbol{v} &= 0, \\ \vdots \quad &\vdots \quad \vdots \\ \boldsymbol{b}_{d_u+1}^\top \boldsymbol{v} &= 0. \end{aligned}$$

Because all hyperplanes are in general position, the intersection of the first d_u hyperplanes constructed from the normal vectors $\boldsymbol{b}_1^\top, \boldsymbol{b}_2^\top, \ldots, \boldsymbol{b}_{d_u}^\top$ will form a linear subspace of dimension d_θ. Thus, we can solve the first d_u equations above for $\boldsymbol{u}^\top = (u_1, u_2, \ldots, u_{d_u})$, expressed parametrically in terms of θ, and can substitute this solution into the last equation $\boldsymbol{b}_{d_u+1}^\top \boldsymbol{v} = 0$. After grouping terms, this substitution yields a single equation of the form $\boldsymbol{a}^\top \boldsymbol{v} = 0$ where $\boldsymbol{a}^\top$ has a zero in its first d_u entries. Thus, $\boldsymbol{a}^\top$ will be collected in the matrix $\boldsymbol{V}$ by Algorithm 3. We now claim that each of the $\binom{n}{d_u+1}$ combinations of $d_u + 1$ hyperplanes yields a unique hyperplane of this form. Suppose not. Suppose $\boldsymbol{B}_1 \neq \boldsymbol{B}_2$ are two matrices with $d_u + 1$ rows selected from $\boldsymbol{A}$, and without loss of generality suppose $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ differ by one row. Following an identical substitution procedure as above, it must be that the rows of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the normal vectors of hyperplanes that share the same intersection. Since $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ differ by one row, this means there is a collection of $d_u + 2$ hyperplanes that intersect in a $d_\theta - 1$ dimensional linear subspace, which contradicts the fact that all hyperplanes in $\boldsymbol{A}$ are in general position. From here we conclude that each of the $\binom{n}{d_u+1}$ combinations of $d_u + 1$ hyperplanes yields a unique hyperplane collected in $\boldsymbol{V}$, so there are at least $\binom{n}{d_u+1}$ rows in $\boldsymbol{V}$.

We now show that Algorithm 3 will return exactly $\binom{n}{d_u+1}$ rows. To do so, we will prove that Algorithm 3 will never discard any rows of $\boldsymbol{A}$ before reaching step 3 with $j = d_u + 1$. By Assumption, none of the hyperplanes in $\mathscr{A}$ are orthogonal to $\mathbb{R}^{d_\theta}$, so no rows are eliminated in step 1. Now for step 2 or 3 with $2 \leq j \leq d_u$, the intersection of j hyperplanes, all in general position, forms a linear subspace of dimension $d - j \geq d_\theta$. Such a linear subspace cannot be contained in a hyperplane that is orthogonal to $\mathbb{R}^{d_\theta}$, since such a hyperplane forms linear subspace of dimension at most $d_\theta - 1$. ■

Proof of Proposition 3.6. Note $\theta \in \partial \mathrm{Proj}_\Theta(\mathcal{C}^*)$ implies that $B_\varepsilon(\theta) \cap \mathrm{Proj}_\Theta(\mathcal{C}^*) \neq \varnothing$ for every $\varepsilon > 0$. Now let $\boldsymbol{s}^*$ denote the sign vector associated with $\mathcal{C}^*$. Furthermore, note that:

$$\theta \in E := \bigcap_{\boldsymbol{s} \in \mathcal{S}(\theta)} \Theta(\boldsymbol{s}).$$

(By convention, if $\mathcal{S}(\theta) = \varnothing$ then $E = \mathbb{R}^{d_\theta}$). Each set $\Theta(\boldsymbol{s})$ is the projection of an open set, and so is open.

As a finite intersection of open sets, E is also open. Thus, we can choose $\varepsilon > 0$ small enough such that $B_\varepsilon(\theta) \subset E$. Since $B_\varepsilon(\theta) \cap \text{Proj}_\Theta(\mathcal{C}^*) \neq \varnothing$ for every $\varepsilon > 0$, this means there exists a $\theta' \in B_\varepsilon(\theta)$ satisfying:

$$\theta' \in \bigcap_{\boldsymbol{s} \in \mathcal{S}(\theta) \cup \{\boldsymbol{s}^*\}} \Theta(\boldsymbol{s}).$$

But then $\mathcal{S}(\theta') \supset \mathcal{S}(\theta)$. ■

Proof of Lemma 3.1. Fix any $\theta \in \Theta$, and note by construction of each cell $\mathcal{C} \in \mathscr{C}$ we have $\boldsymbol{u}, \boldsymbol{u}' \in \mathcal{C}$ implies that $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta)$ for all $\boldsymbol{x} \in \mathcal{X}$. Furthermore, all cells in $\mathscr{C}$ are disjoint, and they partition (3.9). Conclude that each $\mathcal{U}' \in \mathbb{U}$ can be written as a disjoint union of cells in $\mathscr{C}$. The second claim in the Lemma follows immediately from the first claim.

■

Appendix B Auxiliary Results

B.1 Incorporating Additional Assumptions

In this section we will show how the additional assumptions from Section 4 can be accommodated by our theoretical framework. We assume throughout that all assumptions from Section 4 are imposed, which will allow us to demonstrate which components of the main proofs must be revised for each assumption. It is then straightforward to verify that similar definitions and results hold when the researcher instead imposes any combination of the assumptions from Section 4.

Definition B.1. *Suppose Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold. Consider the following conditions:*

(i) For the subset $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ from Assumption 4.1, the equality:

$$P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}) = P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in B \mid \boldsymbol{X} = \boldsymbol{x}'),$$

holds for all $B \in \mathfrak{B}(\mathcal{U})$ and all $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ occurring with positive probability.

(ii) For the elements $U_{k_1}, \ldots, U_{k_Q}$ of the vector $\boldsymbol{U}$ with supports $\mathcal{U}_{k_1}, \ldots, \mathcal{U}_{k_Q}$ the equality:

$$P_{\boldsymbol{U}}(U_{k_q} \leq \tau_{k_q}) = p_{k_q},$$

holds for the pairs (p_{k_q}, τ_{k_q}) from Assumption 4.2, for $q = 1, \ldots, Q$.

(iii) For any $\boldsymbol{x} \in \mathcal{X}$ such that γ is trivial at $\boldsymbol{x} \in \mathcal{X}$, we have:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right),$$

$P_{\boldsymbol{U},\boldsymbol{X}}-a.s$ for all $\boldsymbol{y} \in \mathcal{Y}$.

Then the (joint) identified set $\mathcal{I}^{**}_{\boldsymbol{X}}$ of structures is the set of all pairs $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta)$ satisfying conditions (i), (ii) and (iii) such that there exists a conditional distribution $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ satisfying:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_{B} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{B.1}$$

$P_{\boldsymbol{X}}-$almost surely for every $B \subset \mathcal{Y}$, and such that $P_{\boldsymbol{U}} = P_{\boldsymbol{U}|\boldsymbol{X}} P_{\boldsymbol{X}}$ assigns zero probability to all sets of the form $\{\boldsymbol{u} \in \mathcal{U} : g^{(m_1,m_2)}_{j,k}(A)(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\}$ from Assumption 2.1 and the hyperplanes $\{\boldsymbol{u} \in \mathcal{U} : \langle e_{k_q}, \boldsymbol{u} \rangle = \tau_{k_q}\}$. Furthermore, for a given counterfactual γ, the identified set of counterfactual conditional distributions $\mathcal{P}^{**}_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ is the set of all conditional distributions such that there exists a conditional distribution $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and some pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^{**}_{\boldsymbol{X}}$ defined above satisfying conditions (i), (ii) and (iii), and:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in B \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{U}} \int_{B} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}} \, dP_{\boldsymbol{U}|\boldsymbol{X}}, \tag{B.2}$$

$P_{\boldsymbol{X}}-$almost surely for every $B \subset \mathcal{Y}$. Finally, the identified set for the counterfactual functional $\Psi_{\boldsymbol{x}}$ from (2.9) is the set of all values $\overline{\Psi}_{\boldsymbol{x}}$ such that there exists some $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^{**}_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ satisfying:

$$\overline{\Psi}_{\boldsymbol{x}} := \int_{\mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X}},$$

$P_{\boldsymbol{X}}-$almost surely.

Abusing notation, let $\mathscr{H}$ denote the hyperplane arrangement containing all hyperplanes of the form (2.12) in the main text, and any additional hyperplanes added to implement Assumption 4.2. For Assumption 4.2, we must add all hyperplanes of the form:

$$\{\boldsymbol{u} \in \mathbb{R}^{d_u} : \langle \boldsymbol{e}_{k_q}, \boldsymbol{u} \rangle = \tau_{k_q}\},$$

where $\boldsymbol{e}_{k_q}$ is the vector with 1 in the k^{th}_q entry, and zeroes elsewhere. After including all these additional hyperplanes, redefine $\mathscr{C}$ as the collection of all cells induced by the arrangement $\mathscr{H}$, enumerated as $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_{L'}$.

Corollary B.1. *Suppose Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold, let $v_\ell \in \mathcal{V}_\ell$ be any point from the set $\mathcal{V}_\ell$ and let $\boldsymbol{V}^\dagger$ be a random vector with support on $\mathcal{V}^\dagger := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{L'}\}$. Then $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^{**}_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ if and only if there exists a $\theta \in \Theta$, a collection of finite-dimensional conditional distributions $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$, and finite collections $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}$ that satisfy the conditions:*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{V}^\dagger} \int_{A} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{v}, \theta)\} \, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger} \, dP_{\boldsymbol{V}^\dagger|\boldsymbol{X}}, \tag{B.3}$$

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A \mid \boldsymbol{X} = \boldsymbol{x}) = \int_{\mathcal{V}^\dagger} \int_{A} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger} \, dP_{\boldsymbol{V}^\dagger|\boldsymbol{X}}, \tag{B.4}$$

$P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$, and such that:

(i)

$$P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x}) = P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x}'),$$

for all $\boldsymbol{v} \in \mathcal{V}^\dagger$ *and all* $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ *occurring with positive probability.*

(ii) $P_{\boldsymbol{V}^\dagger}(\boldsymbol{V}^\dagger \in E_{k_q}) = p_{k_q}$ *for* $q = 1, \ldots, Q$, *with* $E_{k_q} := \{\boldsymbol{u} \in \mathbb{R}^{d_u} : \langle \boldsymbol{e}_{k_q}, \boldsymbol{u} \rangle \leq \tau_{k_q}\}$.

(iii) For any $\boldsymbol{x} \in \mathcal{X}$ *such that* γ *is trivial at* $\boldsymbol{x} \in \mathcal{X}$, *we have:*

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell\right) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell\right),$$

$P_{\boldsymbol{V}^\dagger,\boldsymbol{X}}-a.s$ *for all* $\boldsymbol{y} \in \mathcal{Y}$.

Proof of Corollary B.1. Begin by fixing $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^{**}_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$. By Definition B.1, there exists conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and some pair $(P_{\boldsymbol{U}|\boldsymbol{X}}, \theta) \in \mathcal{I}^{**}_{\boldsymbol{X}}$ such that (B.1) and (B.2) are satisfied. Following an identical proof to the proof of Theorem 2.1, construct the distributions $P_{\boldsymbol{Y}|\boldsymbol{V}^\dagger,\boldsymbol{X}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{V}^\dagger,\boldsymbol{X}}$ and $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$ satisfying (B.3) and (B.4) $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$. To complete the proof of necessity, it remains only to show that conditions (i), (ii) and (iii) in the statement of Corollary B.1 are satisfied by the conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}$ and $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$.

By condition (i) in Definition B.1, we have:

$$P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in A \mid \boldsymbol{X} = \boldsymbol{x}) = P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in A \mid \boldsymbol{X} = \boldsymbol{x}'),$$

for all $A \in \mathfrak{B}(\mathcal{U})$ and all $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ occurring with positive probability. Fix any $\boldsymbol{v}_\ell \in \boldsymbol{V}^\dagger$. Then for any $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ we have:

$$\begin{aligned} P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) &= P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in \mathcal{V}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in \mathcal{V}_\ell \mid \boldsymbol{X} = \boldsymbol{x}') \\ &= P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v}_\ell \mid \boldsymbol{X} = \boldsymbol{x}'). \end{aligned}$$

This shows necessity for the independence assumption. Now fix a value $q \in \{1, \ldots, Q\}$. Then condition (ii) in Definition B.1 implies $P_{\boldsymbol{U}}(U_{k_q} \leq \tau_{k_q}) = p_{k_q}$. Recall that:

$$E_{k_q} := \left\{\boldsymbol{u} \in \mathbb{R}^{d_u} : \langle e_{k_q}, \boldsymbol{u} \rangle \leq \tau_{k_q}\right\}.$$

Now note:

$$P_{\boldsymbol{V}^\dagger}(\boldsymbol{V}^\dagger \in E_{k_q}) \overset{(1)}{=} \int_{\mathcal{X}} \int_{E_{k_q}} \int_{\mathcal{Y}} dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \overset{(2)}{=} \int_{E_{k_q}} dP_{\boldsymbol{U}} = p_{k_q}.$$

Here the first equality follows from the definition of $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$, the second equality follows from the fact that $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ integrates to 1 and by the Fubini-Tonelli Theorem, and the equality follows by Assumption 4.2. Since q was arbitrary, this shows necessity for the quantile assumption. Finally, fix any $\boldsymbol{x} \in \mathcal{X}$ such that γ is trivial at $\boldsymbol{x} \in \mathcal{X}$. From condition (iii) in Definition B.1 we have:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}\right),$$

$P_{\boldsymbol{U},\boldsymbol{X}}-$a.s for all $\boldsymbol{y} \in \mathcal{Y}$. This is equivalent to:

$$P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y}, \boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{X} = \boldsymbol{x}\right) = P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}}\left(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{X} = \boldsymbol{x}\right),$$

$P_{\boldsymbol{X}}-$a.s for all $\boldsymbol{y} \in \mathcal{Y}$. Thus we have:

$$\begin{aligned}
P_{\boldsymbol{Y},\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell \mid \boldsymbol{X} = \boldsymbol{x}) &\overset{(1)}{=} \int_{\mathcal{V}_\ell} \int_A dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\overset{(2)}{=} \int_{\mathcal{V}_\ell} P_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{Y} \in A, \boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{X} = \boldsymbol{x})\, d\boldsymbol{u} \\
&\overset{(3)}{=} \int_{\mathcal{V}_\ell} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{X} = \boldsymbol{x})\, d\boldsymbol{u} \\
&\overset{(4)}{=} \int_{\mathcal{V}_\ell} \int_A dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}\, dP_{\boldsymbol{U}|\boldsymbol{X}} \\
&\overset{(5)}{=} P_{\boldsymbol{Y}_\gamma,\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{Y}_\gamma \in A, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

$P_{\boldsymbol{X}}-$a.s., where equalities (1) and (5) follow from the definition of $P_{\boldsymbol{Y},\boldsymbol{V}^\dagger|\boldsymbol{X}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{V}^\dagger|\boldsymbol{X}}$, and equality (3) follows from condition (iii) in Definition B.1. But the previous display is equivalent to:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell),$$

$P_{\boldsymbol{V}^\dagger,\boldsymbol{X}}-$a.s. Thus condition (iii) in the statement of the result is satisfied, and this completes the proof of necessity.

For sufficiency, fix any $\theta \in \Theta$, let $\boldsymbol{V}^\dagger$ be a random vector with support on $\mathcal{V}^\dagger$, and let $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}$, and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}$ be any collections that satisfy conditions (i), (ii) and (iii) and also satisfy conditions (B.3) and (B.4) $P_{\boldsymbol{X}}-$almost surely for all $A \subset \mathcal{Y}$. Following an identical proof to the proof of Theorem 2.1, it is possible to show that the joint distribution constructed from $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}$ and $P_{\boldsymbol{X}}$, and the joint distribution constructed from $P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}$ and $P_{\boldsymbol{X}}$ can be extended to (not necessarily unique) Borel measures $P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{U}}$ on $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$ such that the associated conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{U}|\boldsymbol{X}}$ satisfy (B.1) and (B.2), and such that $P_{\boldsymbol{U}} = P_{\boldsymbol{U}|\boldsymbol{X}} P_{\boldsymbol{X}}$ assigns zero probability to all sets of the form $\{\boldsymbol{u} \in \mathcal{U} : g_{j,k}^{(m_1,m_2)}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = 0\}$ from Assumption 2.1, as well as any added hyperplanes. It remains only to show that these extended conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{U}|\boldsymbol{X}}$ also satisfy conditions (i), (ii) and (iii) from Definition B.1. By assumption we have:

$$P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x}) = P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x}'),$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ and $\boldsymbol{v} \in \mathcal{V}^\dagger$. For every $\boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}$ occurring with positive probability, we have by construction of the extended Borel measure (see the proof of Theorem 2.1):

$$\begin{aligned}
P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in A \mid \boldsymbol{X} = \boldsymbol{x}) &= P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger \in A \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= \sum_{\boldsymbol{v} \in A} P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x})
\end{aligned}$$

$$= \sum_{\boldsymbol{v} \in A} P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger = \boldsymbol{v} \mid \boldsymbol{X} = \boldsymbol{x}')$$
$$= P_{\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{V}^\dagger \in A \mid \boldsymbol{X} = \boldsymbol{x}')$$
$$= P_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{U} \in A \mid \boldsymbol{X} = \boldsymbol{x}'),$$

for any $A \in \mathfrak{B}(\mathcal{U})$, where the third equality follows from the fact that condition (i) is satisfied. Furthermore, for any $q \in \{1, \ldots, Q\}$, by construction of the extended Borel measure (see the proof of Theorem 2.1) we have:

$$P_{\boldsymbol{U}}(U_{k_q} \leq \tau_{k_q}) = P_{\boldsymbol{V}^\dagger}(\boldsymbol{V}^\dagger \in E_{k_q}) = p_{k_q},$$

which follows since condition (ii) is satisfied. This shows that the conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$, $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{U}|\boldsymbol{X}}$ are consistent with condition (ii) in Definition B.1. Finally, fix any $\boldsymbol{x} \in \mathcal{X}$ such that γ is trivial at $\boldsymbol{x} \in \mathcal{X}$. By condition (iii) in the statement of the result, we have:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}\left(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell\right) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell\right),$$

$P_{\boldsymbol{V}^\dagger,\boldsymbol{X}}-$a.s for all $\boldsymbol{y} \in \mathcal{Y}$. This is equivalent to:

$$P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}) = P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y}_\gamma = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}),$$

for all $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}) \in \mathcal{Y} \times \mathcal{X} \times \mathcal{V}^\dagger$. But for any $B \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{U})$:

$$\begin{aligned}
P_{\boldsymbol{Y},\boldsymbol{U},\boldsymbol{X}}((\boldsymbol{Y},\boldsymbol{U}) \in B, \boldsymbol{X} = \boldsymbol{x}) &\overset{(1)}{=} \int_{\mathcal{X}} \int_{\mathcal{V}^\dagger} \int_{\mathcal{Y}} \mathbb{1}\{(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{x}) \in B \times \{\boldsymbol{x}\}\}\, dP_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{V}^\dagger}\, dP_{\boldsymbol{V}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \\
&\overset{(2)}{=} \sum_{(\boldsymbol{y},\boldsymbol{v}) \in \mathcal{Y} \times \mathcal{V}^\dagger} P_{\boldsymbol{Y},\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}) \\
&\overset{(3)}{=} \sum_{(\boldsymbol{y},\boldsymbol{v}) \in \mathcal{Y} \times \mathcal{V}^\dagger} P_{\boldsymbol{Y}_\gamma,\boldsymbol{X},\boldsymbol{V}^\dagger}(\boldsymbol{Y}_\gamma = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{V}^\dagger = \boldsymbol{v}) \\
&\overset{(4)}{=} \int_{\mathcal{X}} \int_{\mathcal{V}^\dagger} \int_{\mathcal{Y}} \mathbb{1}\{(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{x}) \in B \times \{\boldsymbol{x}\}\}\, dP_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{V}^\dagger}\, dP_{\boldsymbol{V}^\dagger|\boldsymbol{X}}\, dP_{\boldsymbol{X}} \\
&\overset{(5)}{=} P_{\boldsymbol{Y}_\gamma,\boldsymbol{U},\boldsymbol{X}}((\boldsymbol{Y}_\gamma,\boldsymbol{U}) \in B, \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

where equalities (1) and (2) follow from the extension in the proof of Theorem 2.1, and equality (3) follows from condition (iii) in the statement of the result. But the previous display is equivalent to:

$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}(\boldsymbol{Y}_\gamma = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}) = P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}),$$

$P_{\boldsymbol{U},\boldsymbol{X}}-$a.s. This completes the proof of sufficiency. ■

To introduce the extension of Theorem 2.2 under the additional Assumptions 4.1, 4.2 and 4.3, we will first describe how to convert these assumptions into the appropriate constraints on the parameter vector $\boldsymbol{\pi}$ in the optimization problems from Theorem 2.2. Let us re-define the optimizing variable $\boldsymbol{\pi}$ as a vector in

$\mathbb{R}^{d_\pi}$ with typical element $\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell)$ representing the probability:

$$P_{\boldsymbol{Y},\boldsymbol{Y}_\gamma,\boldsymbol{V}^\dagger|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Y}_\gamma = \boldsymbol{y}^\star, \boldsymbol{V}^\dagger = \boldsymbol{v}_\ell \mid \boldsymbol{X} = \boldsymbol{x}).$$

Assume (without loss of generality) that each $\boldsymbol{x} \in \mathcal{X}$ is assigned positive probability by the observed distribution. Similar to the constraints (2.15), (2.16), and (2.17) in Theorem 2.2, we must impose the constraints:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) = \sum_{\ell=1}^{L'} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{v}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell), \quad \forall (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}, \tag{B.5}$$

as well as the non-negativity constraints:

$$\pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) \geq 0, \quad \forall (\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{V}^\dagger \times \mathcal{X}, \tag{B.6}$$

and the adding up constraints:

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\ell=1}^{L'} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) = 1, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{B.7}$$

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\ell=1}^{L'} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{v}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) = 1, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{B.8}$$

The independence constraints can be imposed as:

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \pi_{\boldsymbol{x}'}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \widetilde{\mathcal{X}}, \tag{B.9}$$

for $\ell = 1, \ldots, L'$. The quantile constraints can be written as:

$$\sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{y} \in \mathcal{Y}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\boldsymbol{v}_\ell \in \mathcal{V}^\dagger \cap E_{k_q}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell) P_{\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}) = p_{k_q}, \tag{B.10}$$

for $q = 1, \ldots, Q$. Finally, for any $\boldsymbol{x} \in \mathcal{X}$ such that γ is trivial at $\boldsymbol{x} \in \mathcal{X}$, the equilibrium selection constraints can be written as:

$$\sum_{\boldsymbol{y}' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}', \boldsymbol{y}, \boldsymbol{v}_\ell) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}', \boldsymbol{v}_\ell), \tag{B.11}$$

for all $\boldsymbol{y} \in \mathcal{Y}$ and all $\boldsymbol{x} \in \mathcal{X}$ satisfying $c_x(\boldsymbol{x}) = \boldsymbol{x}$.

Corollary B.2. *Under Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3, the identified set for the counterfactual functional $\Psi_{\boldsymbol{x}}$ is given by:*

$$\bigcup_{\theta \in \Theta} [\Psi_{\boldsymbol{x}}^{\ell b}(\theta), \Psi_{\boldsymbol{x}}^{ub}(\theta)],$$

where $\Psi^{\ell b}_{\boldsymbol{x}}(\theta)$ *and* $\Psi^{ub}_{\boldsymbol{x}}(\theta)$ *are determined by the optimization problems:*

$$\Psi^{\ell b}_{\boldsymbol{x}}(\theta) := \min_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L'} \sum_{\boldsymbol{y} \in \mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{v}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell),$$

$$\text{s.t. (B.5), (B.6), (B.7), (B.8), (B.9), (B.10) and (B.11),}$$

and:

$$\Psi^{ub}_{\boldsymbol{y},\boldsymbol{x}}(\theta) := \max_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L'} \sum_{\boldsymbol{y} \in \mathcal{Y}} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star) \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{v}_\ell, \theta)\} \pi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{y}^\star, \boldsymbol{v}_\ell),$$

$$\text{s.t. (B.5), (B.6), (B.7), (B.8), (B.9), (B.10) and (B.11).}$$

The proof of Corollary B.2 follows a similar approach to the previous proofs, and so is omitted.

B.2 Measurability Results

Definition B.2 (Effros-Measurability, Random Set, Selections)**.** *Let* $(\Omega, \mathfrak{A}, P)$ *be a probability space, let* $\mathcal{V}$ *be a Polish space, and let* $\mathcal{O}_\mathcal{V}$ *and* $\mathcal{F}_\mathcal{V}$ *denote the collection of all open and closed sets on* $\mathcal{V}$*. A multifunction* $\mathbb{V} : \Omega \to \mathcal{F}_\mathcal{V}$ *is called Effros-measurable if for every* $A \in \mathcal{O}_\mathcal{V}$ *we have* $\mathbb{V}^-(A) := \{\omega \in \Omega : \mathbb{V}(\omega) \cap A \neq \varnothing\} \in \mathfrak{A}$*. A random closed set is an Effros-measurable closed multifunction defined on a probability space. A random element* $V : \Omega \to \mathcal{V}$ *is called a (measurable) selection of* $\mathbb{V}$ *if* $V(\omega) \in \mathbb{V}(\omega)$ *for* $P-$*almost all* $\omega \in \Omega$*.*

Lemma B.1. *Suppose Assumptions 2.1 and 2.2 hold. Then for each* $\theta \in \Theta$ *and any* $\gamma \in \Gamma$*:*

(i) $(\boldsymbol{x}, \boldsymbol{u}) \mapsto \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)$ *is a random closed set.*

(ii) $(\boldsymbol{x}, \boldsymbol{u}) \mapsto \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)$ *is a random closed set.*

(iii) $\mathrm{Graph}(\mathcal{Y}(\,\cdot\,, \theta)) := \{(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) : \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$*.*

(iv) $\mathrm{Graph}(\mathcal{Y}_\gamma(\,\cdot\,, \theta)) := \{(\boldsymbol{y}^\star, \boldsymbol{x}, \boldsymbol{u}) : \boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\} \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{U})$*.*

(v) $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$ *is measurable.*

(vi) $(\boldsymbol{y}^\star, \boldsymbol{x}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)\}$ *is measurable.*

Proof of Lemma B.1. Clearly $(iii) \implies (v)$ and $(iv) \implies (vi)$. Furthermore, by Theorem 1.3.3 in Molchanov (2017), $(i) \implies (iii)$ and $(ii) \implies (iv)$. Thus, it suffices to prove only (i) and (ii). Since the proof of (ii) is nearly identical to the proof of (i), we will focus on proving (i). To this end, recall from (2.1):

$$\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) := \bigcap_{m_1 \in \mathcal{M}_1} \bigcup_{m_2 \in \mathcal{M}_2} \mathcal{Y}^{(m_1, m_2)}(\boldsymbol{x}, \boldsymbol{u}, \theta),$$

where:

$$\mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x},\boldsymbol{u},\theta) = \left\{\boldsymbol{y} \in \overline{\mathcal{Y}}_{(m_1,m_2)} : \boldsymbol{y} - \varphi^{(m_1,m_2)}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta) = \boldsymbol{0}\right\},$$

Since $\mathcal{Y}$ is finite and equipped with the discrete topology, the function $\boldsymbol{y} \mapsto \boldsymbol{y} - \varphi^{(m_1,m_2)}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is trivially continuous for every tuple $(\boldsymbol{x},\boldsymbol{u},\theta)$ (every function on $\mathcal{Y}$ is continuous when $\mathcal{Y}$ has the discrete topology). Furthermore, under Assumption 2.1 each function $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is continuous in $\boldsymbol{u}$ and measurable in $(\boldsymbol{y},\boldsymbol{x})$ for each θ, implying it is a Caratheodory function. By Lemma 4.51 in Aliprantis and Border (2006), this implies that each function $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is jointly measurable in $(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u})$ for each θ. Since the function $\boldsymbol{y} - \varphi^{(m_1,m_2)}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is a Borel measurable function of the functions $g_{j,k}^{(m_1,m_2)}(A)(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$, it follows that $\boldsymbol{y} - \varphi^{(m_1,m_2)}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is jointly measurable in $(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u})$, and thus is also measurable in $(\boldsymbol{x},\boldsymbol{u})$. Thus, for each θ the function $\boldsymbol{y} - \varphi^{(m_1,m_2)}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{u},\theta)$ is (trivially) continuous in $\boldsymbol{y}$ and measurable in $(\boldsymbol{x},\boldsymbol{u})$, implying it is also a Caratheodory function. Since $\overline{\mathcal{Y}}_{(m_1,m_2)} \subseteq \mathcal{Y}$ is also compact, we have by Corollary 18.8 in Aliprantis and Border (2006) that $\mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x},\boldsymbol{u},\theta)$ is weakly measurable. By Lemma 18.4(1) in Aliprantis and Border (2006) this also implies that $\bigcup_{m_2 \in \mathcal{M}_2} \mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x},\boldsymbol{u},\theta)$ is weakly measurable. By Lemma 18.4(3) in Aliprantis and Border (2006), this in turn implies that $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ is weakly measurable. Since each set $\mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x},\boldsymbol{u},\theta)$ is closed, and since $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ is constructed from $\mathcal{Y}^{(m_1,m_2)}(\boldsymbol{x},\boldsymbol{u},\theta)$ via finite intersections and unions, $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ is also closed. This implies that $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ is an Effros measurable closed-valued multifunction, and thus a random closed set (c.f. Molchanov (2017) p. 72).

■

Appendix C Simplifications when the Model is Complete

Both Theorems 2.1 and 2.2 can be greatly simplified when the researcher's model is complete. In particular, suppose (for this subsection only) that the researchers model is complete, so that for each $\theta \in \Theta$ both $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)$ are either empty, or are singletons. By definition, both of the conditional distributions $P_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{U}}$ and $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X},\boldsymbol{U}}$ are degenerate when the model is complete. This leads us to the following simplification of Theorem 2.1.

Corollary C.1. *Suppose Assumptions 2.1 and 2.2 hold, and that both $\mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta)$ are either empty or are singletons for every $(\boldsymbol{x},\boldsymbol{u})$. Furthermore, let $\boldsymbol{u}_\ell \in \mathcal{U}_\ell$ be any point from the set $\mathcal{U}_\ell$ belonging to the collection $\mathscr{C}$, and let $\boldsymbol{U}^\dagger$ be a random vector with support on $\mathcal{U}^\dagger := \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_L\}$. Then $P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}} \in \mathcal{P}^*_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}$ if and only if there exists a $\theta \in \Theta$ and a collection of finite-dimensional conditional distributions $P_{\boldsymbol{U}^\dagger|\boldsymbol{X}}$ that satisfy the conditions:*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} \in A \mid \boldsymbol{X} = \boldsymbol{x}\right) = \int_{\mathcal{U}^\dagger} \mathbb{1}\{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x},\boldsymbol{u},\theta) \cap A\}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}},$$
$$P_{\boldsymbol{Y}_\gamma|\boldsymbol{X}}\left(\boldsymbol{Y}_\gamma \in A \mid \boldsymbol{X} = \boldsymbol{x}\right) = \int_{\mathcal{U}^\dagger} \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}_\gamma(\boldsymbol{x},\boldsymbol{u},\theta) \cap A\}\, dP_{\boldsymbol{U}^\dagger|\boldsymbol{X}},$$

$P_{\boldsymbol{X}}-$*almost surely for all* $A \subset \mathcal{Y}$.

This simplified version of Theorem 2.1 in turn leads to a simplification of Theorem 2.2. In particular, enumerate the points in $\mathcal{U}^\dagger$ from Corollary C.1 as $\{\boldsymbol{u}_\ell\}_{\ell=1}^L$, and let $\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}$ be a vector with typical element $\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell)$. Assume for simplicity that all $\boldsymbol{x} \in \mathcal{X}$ are assigned positive probability. Now consider the set of constraints:

$$P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}\right) = \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y} = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell), \tag{C.1}$$

for each $\boldsymbol{y}$ and for all $\boldsymbol{x}$. In addition, we must impose the non-negativity constraints:

$$\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell) \geq 0, \tag{C.2}$$

for $\ell = 1, \ldots, L$ and for all $\boldsymbol{x} \in \mathcal{X}$, and the adding up conditions:

$$\sum_{\ell=1}^{L} \pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell) = 1, \qquad \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \mathbb{1}\{\boldsymbol{y}^\star = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell) = 1, \tag{C.3}$$

for all $\boldsymbol{x} \in \mathcal{X}$. The constraints (C.1) - (C.3) in the complete case should be compared to the constraints (2.15) - (2.17) in the general case. We then have the following simplification of Theorem 2.2.

Corollary C.2. *Suppose Assumptions 2.1 and 2.2 hold, and that both* $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)$ *and* $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta)$ *are either empty, or are singletons for each* $\theta \in \Theta$. *Then the identified set for the counterfactual functional* $\Psi_{\boldsymbol{x}}$ *is given by:*

$$\bigcup_{\theta \in \Theta} [\Psi_{\boldsymbol{x}}^{\ell b}(\theta), \Psi_{\boldsymbol{x}}^{ub}(\theta)],$$

where $\Psi_{\boldsymbol{x}}^{\ell b}(\theta)$ *and* $\Psi_{\boldsymbol{x}}^{ub}(\theta)$ *are determined by the optimization problems:*

$$\Psi_{\boldsymbol{x}}^{\ell b}(\theta) := \min_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\mathbb{1}\{\boldsymbol{y}^\star = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell), \ \textit{s.t.} \ \text{(C.1)}, \ \text{(C.2)} \ \textit{and} \ \text{(C.3)},$$

and:

$$\Psi_{\boldsymbol{x}}^{ub}(\theta) := \max_{\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}} \sum_{\boldsymbol{y}^\star \in \mathcal{Y}} \sum_{\ell=1}^{L} \psi_{\boldsymbol{x}}(\boldsymbol{y}^\star)\mathbb{1}\{\boldsymbol{y}^\star = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}_\ell, \theta)\}\pi_{\boldsymbol{x}}(\boldsymbol{u}_\ell), \ \textit{s.t.} \ \text{(C.1)}, \ \text{(C.2)} \ \textit{and} \ \text{(C.3)}.$$

Following the discussion in Appendix B.1, it is also straightforward to show that this result can be extended to accommodate the additional assumptions from Section 4.

Appendix D Details on the Minimal Relevant Partition

D.1 Uniqueness of the Minimal Relevant Partition

Consider the collections:

$$\mathscr{U} := \{\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta) : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}\}, \tag{D.1}$$

$$\mathscr{U}_\gamma := \{\mathcal{U}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \theta) : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}\}. \tag{D.2}$$

where $\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and $\mathcal{U}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \theta)$ are the sets defined in (3.7) and (3.8). Let $\pi(\mathscr{U}, \mathscr{U}_\gamma)$ denote the $\pi-$system generated by the sets in $\mathscr{U}$ and $\mathscr{U}_\gamma$.[48] Furthermore, let $\mathbb{U}$ denote the finest partition of (3.9) that can be formed by the sets in $\pi(\mathscr{U}, \mathscr{U}_\gamma)$.[49] Following convention, we assume $\varnothing \notin \mathbb{U}$. It is not difficult to see that each set in $\mathbb{U}$ is of the form:

$$\left(\bigcap_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y}^\dagger \times \mathcal{X}^\dagger} \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta) \right) \cap \left(\bigcap_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y}^{\dagger\dagger} \times \mathcal{X}^{\dagger\dagger}} \mathcal{U}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \theta) \right),$$

for various subsets $\mathcal{Y}^\dagger, \mathcal{Y}^{\dagger\dagger} \subset \mathcal{Y}$ and $\mathcal{X}^\dagger, \mathcal{X}^{\dagger\dagger} \subset \mathcal{X}$. This intuition may help the reader to understand the results that follow.

Lemma D.1. *Suppose Assumptions 2.1 and 2.2 hold, and suppose at least one set of the form (3.7) and (3.8) is nonempty. Then $\mathbb{U}$ is the unique MRP.*

Proof of Lemma D.1. First we verify that $\mathbb{U}$ satisfies the conditions in Definition 3.5, so that it is an MRP. Since $\mathbb{U}$ is a partition of (3.9) by construction, we need only verify that $\mathbb{U}$ satisfies part (ii) of Definition 3.5. There are two ways that part (ii) of Definition 3.5 could be violated by $\mathbb{U}$:

(a) (Redundancy) There exists two unique subsets $\mathcal{U}', \mathcal{U}'' \in \mathbb{U}$ and elements $\boldsymbol{u}' \in \mathcal{U}'$ and $\boldsymbol{u}'' \in \mathcal{U}''$ such that $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta) = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta) = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ for all $\boldsymbol{x} \in \mathcal{X}$.

(b) (Oversized) There exists two elements $\boldsymbol{u}$ and $\boldsymbol{u}'$ from (3.9) that belong to the same set $\mathcal{U}' \in \mathbb{U}$, but either $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta) \neq \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}, \theta) \neq \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta)$ for some $\boldsymbol{x} \in \mathcal{X}$.

Suppose instead $\mathbb{U}$ satisfies (a). Then there exists two unique subsets $\mathcal{U}', \mathcal{U}'' \in \mathbb{U}$ and elements $\boldsymbol{u}' \in \mathcal{U}'$ and $\boldsymbol{u}'' \in \mathcal{U}''$ such that $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta) = \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ and $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta) = \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ for all $\boldsymbol{x} \in \mathcal{X}$. By construction, all sets in $\mathbb{U}$ can be written as nonempty intersections of sets in $\mathscr{U}$ and $\mathscr{U}_\gamma$ or their complements. Thus, since $\mathcal{U}', \mathcal{U}'' \in \mathbb{U}$ are unique sets, and $\boldsymbol{u}' \in \mathcal{U}'$ and $\boldsymbol{u}'' \in \mathcal{U}''$, it must be that there exists at least one set (or its complement) from $\mathscr{U}$ or $\mathscr{U}_\gamma$ used in the intersection defining $\mathcal{U}'$ that contains $\boldsymbol{u}'$ as an element, but does not contain $\boldsymbol{u}''$. Without loss of generality, suppose that $\boldsymbol{u}' \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ (for some $(\boldsymbol{y}, \boldsymbol{x})$) but

[48] Recall that a $\pi-$system is a collection of sets that is closed under intersections and complements.

[49] For two partitions $\mathscr{P}_1$ and $\mathscr{P}_2$ of a set A, we say that $\mathscr{P}_1$ is finer than $\mathscr{P}_2$ if for every $B \in \mathscr{P}_1$ there exists a $C \in \mathscr{P}_2$ such that $B \subset C$. This relation generates a partial ordering over all possible partitions of a set. Since we restrict attention to partitions of (3.9) using the finite collection of sets in $\mathscr{U}$ and $\mathscr{U}_\gamma$, and since every finite partially order set has a minimal and maximal element, our finest partition is well-defined.

$\boldsymbol{u}'' \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. This implies that $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{u}', \boldsymbol{x}, \theta)$ but $\boldsymbol{y} \notin \mathcal{Y}(\boldsymbol{u}'', \boldsymbol{x}, \theta)$. But this contradicts the fact that $\mathcal{Y}(\boldsymbol{u}', \boldsymbol{x}, \theta) = \mathcal{Y}(\boldsymbol{u}'', \boldsymbol{x}, \theta)$ for all $\boldsymbol{x} \in \mathcal{X}$. Conclude that $\mathbb{U}$ satisfies $\neg(a)$.

Now suppose instead that $\mathbb{U}$ satisfies (b). Then there exists a subset $\mathcal{U}' \in \mathbb{U}$ and elements $\boldsymbol{u}', \boldsymbol{u}'' \in \mathcal{U}'$ such that either $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta) \neq \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ or $\mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}', \theta) \neq \mathcal{Y}_\gamma(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ for some $\boldsymbol{x} \in \mathcal{X}$. Without loss of generality, suppose $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta) \neq \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ for some $\boldsymbol{x} \in \mathcal{X}$. Then there is some $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ or $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}'', \theta)$ such that either $\boldsymbol{u}' \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and $\boldsymbol{u}'' \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ or $\boldsymbol{u}' \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and $\boldsymbol{u}'' \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. Without loss of generality, suppose $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}', \theta)$ is such that $\boldsymbol{u}' \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and $\boldsymbol{u}'' \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. Since $\mathcal{U}' \in \mathbb{U}$, it must be that $\mathcal{U}'$ can be written as the intersection of sets in $\mathscr{U}$ and $\mathscr{U}_\gamma$ or their complements. Since $\boldsymbol{u}' \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and $\boldsymbol{u}' \in \mathcal{U}'$, it must be that $\mathcal{U}' \subset \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. But this contradicts the fact that $\boldsymbol{u}'' \in \mathcal{U}'$ and $\boldsymbol{u}'' \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$. Conclude that $\mathbb{U}$ satisfies $\neg(b)$. Since $\mathbb{U}$ satisfies $\neg(a) \wedge \neg(b)$, $\mathbb{U}$ cannot violate condition (ii) in Definition 3.5. This verifies that $\mathbb{U}$ is an MRP.

Now we prove that $\mathbb{U}$ is unique. To this end, let $\mathbb{U}' \neq \mathbb{U}$ be any other partition of (3.9). It suffices to show that $\mathbb{U}'$ cannot satisfy part (ii) of Definition 3.5. To show this, we will show that $\mathbb{U}'$ must satisfy either (a) or (b) above. First, if there exists $\mathcal{U}' \in \mathbb{U}'$ such that $\mathcal{U}' \subset \mathcal{U}''$ for some $\mathcal{U}'' \in \mathbb{U}$, then $\mathbb{U}'$ satisfies (a); otherwise, this contradicts the fact that $\mathbb{U}$ satisfies $\neg(b)$. Similarly, if there exists $\mathcal{U}' \in \mathbb{U}'$ such that $\mathcal{U}'' \subset \mathcal{U}'$ for some $\mathcal{U}'' \in \mathbb{U}$, then $\mathbb{U}'$ satisfies (b); otherwise, this contradicts the fact that $\mathbb{U}$ satisfies $\neg(a)$. Finally, suppose there exists $\mathcal{U}' \in \mathbb{U}'$ such that $\mathcal{U}' \cap \mathcal{U}'' \neq \varnothing$ and $\mathcal{U}' \cap \mathcal{U}''' \neq \varnothing$ for two disjoint $\mathcal{U}'', \mathcal{U}''' \in \mathbb{U}$. Then since $\mathbb{U}$ satisfies $\neg(a)$, it must be that $\mathbb{U}'$ satisfies (b). Since $\mathbb{U}'$ forms a partition of (3.9) by assumption, this exhausts all cases and completes the proof.

■

D.2 Relation to Core Determining Classes

In the previous subsection we showed the MRP is unique. In this section we relate the MRP to core-determining classes for Artstein's inequalities. To place the MRP and Artstein's inequalities on an equal footing, throughout this section we consider only the case where γ is trivial at all values of $\boldsymbol{x} \in \mathcal{X}$ (see Definition 4.1). This condition implies that the collection (D.2) is contained in the collection (D.1), and is most relevant to the case when the researcher wishes to use Artstein's inequalities or the MRP to learn about structural parameters rather than counterfactuals. This motivation aligns with typical applications of Artstein's inequalities to date. However, any other assumption that ensures the collection (D.2) is contained in the collection (D.1) can also be used; for instance, if γ is a counterfactual that modifies only the values of $\boldsymbol{x} \in \mathcal{X}$ through a (nontrivial) function $c_{\boldsymbol{x}} : \mathcal{X} \to \mathcal{X}$, but changes nothing else. The following result is from Artstein (1983).[50]

Theorem D.1 (Artstein (1983))**.** *Let $(\Omega, \mathfrak{A}, P)$ be a probability space, let $\mathbb{V} : \Omega \to \mathcal{F}_{\mathcal{V}}$ be a random closed set on $\mathcal{V} \subset \mathbb{R}^d$ (with the usual topology). Then a (Borel) probability measure P_V on $\mathbb{R}^d$ is the distribution of*

[50] The statement here is taken from Theorem 2.13 in Molchanov and Molinari (2018).

a selection of the random closed set $\mathbb{V}$ *if and only if:*

$$P_V(A) \leq P(\omega : \mathbb{V}(\omega) \cap A \neq \varnothing),$$

for all open $A \subseteq \mathcal{V}$.

Using the notation in Theorem D.1, a collection of open sets $\mathcal{O}$ is called a core-determining class for the random set $\mathbb{V}$ if a Borel probability measure Q on $\mathbb{R}^d$ satisfies the inequalities:

$$Q(A) \leq P(\omega : \mathbb{V}(\omega) \cap A \neq \varnothing),$$

for all open $A \subseteq \mathcal{V}$ if and only if it satisfies the inequalities:

$$Q(A) \leq P(\omega : \mathbb{V}(\omega) \cap A \neq \varnothing),$$

for all $A \in \mathcal{O}$.[51] Core determining classes are not unique, but are useful to reduce the number of inequality constraints needed to characterize the distribution of a selection from a random set. The sets in a core-determining class are called test sets, and we say that one core-determining class $\mathcal{K}$ is smaller than another $\mathcal{K}'$ if the number of test sets in $\mathcal{K}$ is smaller than the number of test sets in $\mathcal{K}'$. To the best of our knowledge, a general characterization of the smallest core-determining class remains an open research question.

The following example shows that the MRP is generally not a core-determining class.

Example 3. *Let* $\mathcal{Y} = \{-1, 0, 1\}$, $\mathcal{U} = \mathbb{R}^2$, *and for* $\boldsymbol{u} \in \mathcal{U}$ *consider the set* $\mathcal{Y}(\boldsymbol{u}) = \mathcal{Y}^{(1)}(\boldsymbol{u}) \cup \mathcal{Y}^{(2)}(\boldsymbol{u})$ *where:*

$$\begin{aligned}
\mathcal{Y}^{(1)}(\boldsymbol{u}) &= \{y \in \{0, 1\} : y = \mathbb{1}\{u_1 \geq 0\}\}, \\
\mathcal{Y}^{(2)}(\boldsymbol{u}) &= \{y \in \{-1\} : y = -\mathbb{1}\{u_1 \geq 0\}\mathbb{1}\{u_2 \geq 0\}\}.
\end{aligned}$$

Furthermore, we define the set:

$$\mathcal{U}(y) := \mathit{int}\{\boldsymbol{u} \in \mathbb{R}^2 : y \in \mathcal{Y}(\boldsymbol{u})\}.$$

By Lemma D.1 the unique MRP is given by $\mathbb{U} = \{A_1, A_2, A_3\}$ *where:*

$$\begin{aligned}
A_1 &= \mathcal{U}(0) \cap \mathcal{U}(1)^c \cap \mathcal{U}(-1)^c = \{\boldsymbol{u} \in \mathbb{R}^2 : u_1 < 0\}, \\
A_2 &= \mathcal{U}(0)^c \cap \mathcal{U}(1) \cap \mathcal{U}(-1) = \{\boldsymbol{u} \in \mathbb{R}^2 : u_1 > 0,\ u_2 > 0\}, \\
A_3 &= \mathcal{U}(0)^c \cap \mathcal{U}(1) \cap \mathcal{U}(-1)^c = \{\boldsymbol{u} \in \mathbb{R}^2 : u_1 > 0,\ u_2 < 0\}.
\end{aligned}$$

Note all other intersections of the sets $\mathcal{U}(0)$, $\mathcal{U}(1)$, $\mathcal{U}(-1)$ *or their complements are empty. Consider the random closed set* $cl(\mathcal{U}(Y))$, *and consider the set of selectionable distributions* $P_{\boldsymbol{U}}$ *that assign zero probability to the set* $\{\boldsymbol{u} \in \mathbb{R}^2 : u_1 = 0$ *or* $u_2 = 0\}$. *Suppose that* $P_Y(Y = 0) = 0.50$ *and* $P_Y(Y = 1) = P_Y(Y = -1) =$

[51] C.f. Definition 2.20 in Molchanov and Molinari (2018).

0.25. *Using the sets in* $\mathbb{U}$ *as test sets for Artstein's inequalities, we have:*

$$P_{\boldsymbol{U}}(A_1) \leq P_Y(cl(\mathcal{U}(Y)) \cap A_1 \neq \varnothing) = P_Y(Y = 0) = 0.50, \tag{D.3}$$

$$P_{\boldsymbol{U}}(A_2) \leq P_Y(cl(\mathcal{U}(Y)) \cap A_2 \neq \varnothing) = P_Y(Y = 1) + P_Y(Y = -1) = 0.50, \tag{D.4}$$

$$P_{\boldsymbol{U}}(A_3) \leq P_Y(cl(\mathcal{U}(Y)) \cap A_3 \neq \varnothing) = P_Y(Y = 1) = 0.25. \tag{D.5}$$

Now consider the inequality:

$$P_{\boldsymbol{U}}(A_2 \cup A_3) \leq P_Y(cl(\mathcal{U}(Y)) \cap (A_2 \cup A_3) \neq \varnothing) = P_Y(Y = 1) + P_Y(Y = -1) = 0.50.$$

Since A_2 *and* A_3 *are open,* $A_2 \cup A_3$ *is also open. Thus, the inequality in the previous display is included in Artstein's inequalities, but is not implied by any combination of* (D.3)*,* (D.4) *or* (D.5)*. Conclude that the test sets from the MRP do not form a core-determining class.*

Remark D.1. *The fact that the MRP is not generally a core-determining class does not affect our identification results, since we do not rely on a characterization of the identified set based on Artstein's inequalities.*

The counterexample uses the fact that the MRP is not closed under unions in order to show that it is not a core-determining class when the model is incomplete. Although the MRP is not a core-determining class in the general case, there is a connection between the MRP and core-determining classes. The following proposition shows that the MRP is a core-determining class when the model is complete and coherent. This occurs, for instance, in Tebaldi et al. (2021) where the concept of the MRP was first introduced.

Proposition D.1. *Fix* $\theta \in \Theta$*, suppose that Assumptions 2.1 and 2.2 hold, suppose that* γ *is trivial at all* $\boldsymbol{x} \in \mathcal{X}$*, and suppose that* $P_{\boldsymbol{U}|\boldsymbol{X}}$ *assigns zero probability to* $\mathcal{H}(\theta)$ *from* (3.6)*. Furthermore, suppose that* $\mathcal{Y}(\boldsymbol{x}, \boldsymbol{u}, \theta)$ *is nonempty and a singleton for every* $(\boldsymbol{x}, \boldsymbol{u})$ *(that is, the model is complete and coherent). Then:*

$$P_{\boldsymbol{U}|\boldsymbol{X}}(A \mid \boldsymbol{X} = \boldsymbol{x}) \leq P_{\boldsymbol{Y}|\boldsymbol{X}}(cl(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap A \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}), \tag{D.6}$$

holds $P_{\boldsymbol{X}}-a.s.$ *for all open* $A \subseteq \mathcal{U}$ *if and only if:*

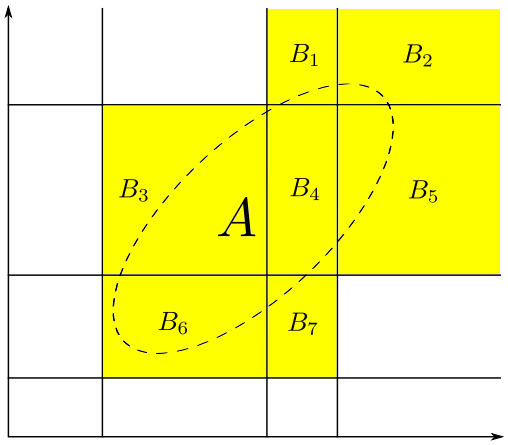
$$P_{\boldsymbol{U}|\boldsymbol{X}}(B \mid \boldsymbol{X} = \boldsymbol{x}) \leq P_{\boldsymbol{Y}|\boldsymbol{X}}(cl(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap B \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}), \tag{D.7}$$

holds for all $B \in \mathbb{U}$*. That is,* $\mathbb{U}$ *is a core-determining class for the random closed set* $cl(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta))$*.*

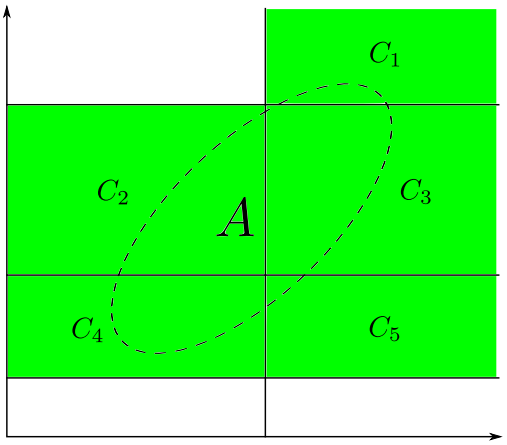
Proof of Proposition D.1. By Lemma 3.1 the sets in $\mathbb{U}$ are disjoint unions of open cells, so all sets in $\mathbb{U}$ are open. Thus, (D.6) implies (D.7). For the reverse, suppose $P_{\boldsymbol{U}|\boldsymbol{X}}$ satisfies (D.7), and suppose by way of contradiction that there exists an open set $A \subset \mathcal{U}$ such that:

$$P_{\boldsymbol{U}|\boldsymbol{X}}(A \mid \boldsymbol{X} = \boldsymbol{x}) > P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap A \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}), \tag{D.8}$$

for some $\boldsymbol{x} \in \mathcal{X}$ assigned positive probability. This is clearly not possible if $A = \varnothing$, so assume that $A \neq \varnothing$.



(a) An arbitrary open set A is covered (modulo boundaries) by the sets $B_1, \ldots, B_7$ from the MRP. $A \cap \overline{\mathcal{U}}$ is exactly the subset of A that does not intersect the boundaries of the hyperplanes.



(b) For each $\boldsymbol{x}$, we can find sets $\{C_m\}_{m=1}^5 \subset \{\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)\}_{\boldsymbol{y} \in \mathcal{Y}}$ to cover the sets $B_1, \ldots, B_7$ from the MRP in Figure 5(a). In a complete model, these sets are disjoint.

Figure 5: An illustration of the proof strategy for Proposition D.1.

Under the conditions in the proposition, the MRP forms a partition of the set:[52]

$$\overline{\mathcal{U}} := \bigcup_{(\boldsymbol{y}, \boldsymbol{x})} \mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta).$$

There must exist a minimal collection of disjoint sets $B_1, \ldots, B_K \in \mathbb{U}$ such that $A \cap \overline{\mathcal{U}} = (A \cap B_1) \cup \ldots \cup (A \cap B_K)$ (see Figure 5 for an illustration). From the discussion in Appendix D.1, each set $B_k \in \mathbb{U}$ is a nonempty intersection of the sets $\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ in $\mathscr{U}(\theta)$. Thus:

$$\begin{aligned}
&\operatorname{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap (A \cap \overline{\mathcal{U}}) \neq \varnothing \\
\iff &\operatorname{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap ((A \cap B_1) \cup \ldots \cup (A \cap B_K)) \neq \varnothing \\
\iff &\operatorname{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap (B_1 \cup \ldots \cup B_K) \neq \varnothing.
\end{aligned}$$

Now note $\overline{\mathcal{U}}^c$ is exactly the subset of $\mathcal{U}$ not covered by the sets $\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ from $\mathscr{U}(\theta)$. When the model is coherent, $\overline{\mathcal{U}}^c$ is $\mathcal{H}(\theta)$, and $P_{\boldsymbol{U}|\boldsymbol{X}}$ assigns zero probability to this set by assumption. Conclude that $P_{\boldsymbol{U}|\boldsymbol{X}}(A \cap \overline{\mathcal{U}}^c \mid \boldsymbol{X} = \boldsymbol{x}) = 0$. Thus from (D.8) and the proof so far, we must have:

$$\begin{aligned}
&P_{\boldsymbol{U}|\boldsymbol{X}}((A \cap B_1) \cup \ldots \cup (A \cap B_K) \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{U}|\boldsymbol{X}}(A \cap \overline{\mathcal{U}} \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{U}|\boldsymbol{X}}(A \mid \boldsymbol{X} = \boldsymbol{x}) \\
&> P_{\boldsymbol{Y}|\boldsymbol{X}}(\operatorname{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap A \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\operatorname{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap (A \cap \overline{\mathcal{U}}) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\operatorname{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap ((A \cap B_1) \cup \ldots \cup (A \cap B_K)) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

[52] In particular when γ is trivial we have $\{\mathcal{U}_\gamma(\boldsymbol{y}, \boldsymbol{x}, \theta)\}_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}} \subseteq \{\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)\}_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}}$.

for some $\boldsymbol{x} \in \mathcal{X}$ assigned positive probability. Let $C_1, \ldots, C_M$ be a collection of sets with $C_m \in \{\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)\}_{\boldsymbol{y} \in \mathcal{Y}}$ such that each C_m contains at least one of the sets $B_1, \ldots, B_K$ (see Figure 5 for an illustration). Since $A \neq \varnothing$, by construction of the sets $B_1, \ldots, B_K \in \mathbb{U}$ and $C_1, \ldots, C_M$ we have:

$$\begin{aligned}
\{\boldsymbol{y} \in \mathcal{Y} : \text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap A \neq \varnothing\} &= \{\boldsymbol{y} \in \mathcal{Y} : \text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap (A \cap B_1) \cup \ldots \cup (A \cap B_K) \neq \varnothing\} \\
&= \{\boldsymbol{y} \in \mathcal{Y} : \text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap (B_1 \cup \ldots \cup B_K) \neq \varnothing\} \\
&= \{\boldsymbol{y} \in \mathcal{Y} : \text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap (C_1 \cup \ldots \cup C_M) \neq \varnothing\} \\
&= \bigcup_{m=1}^{M} \{\boldsymbol{y} \in \mathcal{Y} : \text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)) \cap C_m \neq \varnothing\},
\end{aligned}$$

where the last union is a disjoint union.[53] But since $P_{\boldsymbol{U}|\boldsymbol{X}}$ satisfies (D.7) by assumption, we have:

$$\begin{aligned}
P_{\boldsymbol{U}|\boldsymbol{X}}(A \mid \boldsymbol{X} = \boldsymbol{x}) &= P_{\boldsymbol{U}|\boldsymbol{X}}((A \cap B_1) \cup \ldots \cup (A \cap B_K) \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\leq P_{\boldsymbol{U}|\boldsymbol{X}}(B_1 \cup \ldots \cup B_K \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\leq \sum_{k=1}^{K} P_{\boldsymbol{U}|\boldsymbol{X}}(B_k \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\leq \sum_{k=1}^{K} P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap B_k \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\leq \sum_{m=1}^{M} P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap C_m \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}\left(\bigcup_{m=1}^{M} \{\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap C_m \neq \varnothing\} \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap (C_1 \cup \ldots \cup C_M) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap (B_1 \cup \ldots \cup B_K) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap (A \cap B_1) \cup \ldots \cup (A \cap B_K) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&= P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap (A \cap \overline{\mathcal{U}}) \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}) \\
&\leq P_{\boldsymbol{Y}|\boldsymbol{X}}(\text{cl}(\mathcal{U}(\boldsymbol{Y}, \boldsymbol{X}, \theta)) \cap A \neq \varnothing \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

$P_{\boldsymbol{X}}-$a.s., which clearly contradicts (D.8). ■

For instance, this result implies that the MRP is a core-determining class for Artstein's inequalities when the model is complete and coherent, and the counterfactual under consideration is either trivial—in which case the researcher might use the MRP to learn about structural parameters—or under any other assumption that ensures the collection (D.1) is contained in the collection (D.2). When the model is incomplete, the relation between the MRP and core-determining classes is less clear. We leave a full investigation of this relation to future work.

[53] This follows from the fact that each C_m is a set of the form $\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta)$ and the sets $C_1, \ldots, C_M$ are disjoint, so $\text{cl}(\mathcal{U}(\boldsymbol{y}, \boldsymbol{x}, \theta))$ can "hit" only **one** of the sets in $C_1, \ldots, C_M$ for a fixed $(\boldsymbol{y}, \boldsymbol{x})$. Note this step is not possible without the assumption of model completeness.

Appendix E Consistency

Here we introduce a consistency result from Gu and Russell (2021). This result holds in a more general environment, and subsumes many possible applications of our framework as a special case. In the general environment, the researcher wishes to compute bounds on a functional $\mathbb{E}_P[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]$, where $\psi : \mathcal{W} \times \Delta^{d_\pi} \times \Theta \to \mathbb{R}$, where $\mathcal{W} \subset \mathbb{R}^{d_w}$ denotes the support of the observed random vector $\boldsymbol{W}_i$, and $\Delta^{d_\pi} \times \Theta$ denotes the parameter space with typical elements $(\boldsymbol{\pi}, \theta)$. The values of $(\boldsymbol{\pi}, \theta)$ are constrained by moment inequalities of the form:

$$\mathbb{E}_P[m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] \leq 0, \text{ for } j \in \mathcal{J}(\theta),$$

where $\mathcal{J}(\theta)$ is a finite index set that may depend on θ. Note this does not rule out moment equalities, since each moment equality can be equivalently written as a combination of two moment inequalities. Now define the projection:

$$\Delta^*(\theta, P, b) := \left\{\boldsymbol{\pi} \in \Delta^{d_\pi} : \mathbb{E}_P[m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] \leq b \text{ for } j \in \mathcal{J}(\theta)\right\}.$$

Furthermore, define:

$$\Psi^*(P, b) = \bigcup_{\theta \in \Theta} \left[\min_{\boldsymbol{\pi} \in \Delta^*(\theta, P, b)} \mathbb{E}_P[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)], \max_{\boldsymbol{\pi} \in \Delta^*(\theta, P, b)} \mathbb{E}_P[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]\right].$$

When computing $\Psi^*(P, b)$ both θ and $\boldsymbol{\pi}$ are nuisance parameters, although they enter the bounding problem in different ways. Under Assumption E.1 ahead, the identified set for $\mathbb{E}_P[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]$, denoted by $\Psi^*(P)$, is given by $\Psi^*(P) := \Psi^*(P, 0)$. Gu and Russell (2021) study the consistency properties of the sample analog estimator for this representation of $\Psi^*(P, 0)$, replacing the probability measure P by the empirical measure $\mathbb{P}_n$. In particular, define:

$$\mathbb{E}_{\mathbb{P}_n}[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] := \frac{1}{n}\sum_{i=1}^n \psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta), \qquad \mathbb{E}_{\mathbb{P}_n}[m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] := \frac{1}{n}\sum_{i=1}^n m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta), \text{ for } j \in \mathcal{J}(\theta).$$

Then the proposed sample analog estimator is given by:

$$\Psi^*(\mathbb{P}_n, b_n) = \bigcup_{\theta \in \Theta} \left[\min_{\boldsymbol{\pi} \in \Delta^*(\theta, \mathbb{P}_n, b_n)} \mathbb{E}_{\mathbb{P}_n}[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)], \max_{\boldsymbol{\pi} \in \Delta^*(\theta, \mathbb{P}_n, b_n)} \mathbb{E}_{\mathbb{P}_n}[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]\right],$$

where $b_n \downarrow 0$ is a deterministic sequence described below. Define the sequence $\{\eta_n(\theta)\}_{n=1}^\infty$ as:

$$\eta_n(\theta) := \max\left\{\max_{j \in \mathcal{J}(\theta)} \sup_{\boldsymbol{\pi} \in \Delta^{d_\pi}} |\mathbb{E}_{\mathbb{P}_n}[m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] - \mathbb{E}_P[m_j(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]|, \sup_{\boldsymbol{\pi} \in \Delta^{d_\pi}} |\mathbb{E}_{\mathbb{P}_n}[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)] - \mathbb{E}_P[\psi(\boldsymbol{W}_i, \boldsymbol{\pi}, \theta)]|\right\}.$$

Let $\mathscr{P}$ denote the collection of all probability measures on $\mathcal{W}$. The consistency result depends on the following assumption.

Assumption E.1. *The parameter space $(\Theta, \Delta^{d_\pi}, \mathcal{P})$ satisfies the following: (i) For each $\theta \in \Theta$, the function $\psi(\cdot, \theta) : \mathcal{W} \times \Delta^{d_\pi} \to \mathbb{R}$ is measurable in $\boldsymbol{W}$ and linear in $\boldsymbol{\pi}$ with a (possibly data-dependent) Lipschitz*

constant $C(\theta)$ *satisfying* $\sup_{\theta\in\Theta} C(\theta) < \infty$ *a.s. (ii) For each* $\theta \in \Theta$ *the set* $\mathcal{J}(\theta)$ *is finite, and for each* $j \in \mathcal{J}(\theta)$ *the functions* $m_j(\cdot, \theta) : \mathcal{W} \times \Delta^{d_\pi} \to \mathbb{R}$ *are measurable in* $\boldsymbol{W}$ *and linear in* $\boldsymbol{\pi} \in \Delta^{d_\pi}$*. (iii) For every* $P \in \mathcal{P} \subset \mathscr{P}$ *there exists* $(\boldsymbol{\pi}, \theta) \in \Delta^{d_\pi} \times \Theta$ *such that* $\mathbb{E}_P[m_j(\boldsymbol{W}, \boldsymbol{\pi}, \theta)] \leq 0$ *for all* $j \in \mathcal{J}(\theta)$*. (iv) There exists a finite subset* $\Theta' \subset \Theta$ *such that:*

$$\begin{aligned}\left\{\boldsymbol{\pi} \in \Delta^{d_\pi} : \exists \theta \in \Theta \text{ s.t. } \mathbb{E}_P[m_j(\boldsymbol{W}, \boldsymbol{\pi}, \theta)] \leq 0 \text{ for } j \in \mathcal{J}(\theta)\right\} \\ = \left\{\boldsymbol{\pi} \in \Delta^{d_\pi} : \exists \theta \in \Theta' \text{ s.t. } \mathbb{E}_P[m_j(\boldsymbol{W}, \boldsymbol{\pi}, \theta)] \leq 0 \text{ for } j \in \mathcal{J}(\theta)\right\}.\end{aligned}$$

(v) The researcher has a sample $\{\boldsymbol{W}_i\}_{i=1}^n$*, with each* $\boldsymbol{W}_i$ *an i.i.d. draw from some* $P \in \mathcal{P}$*.*

The components of this assumption are either standard, or are easily verified for a variety of applications of our method. The assumption allows for an objective function and constraints that are data-dependent, so long as they are linear in the parameter vector $\boldsymbol{\pi}$ from Theorem 2.2, and so long as the objective function satisfies a mild Lipschitz continuity condition with respect to θ (see conditions (i) and (ii)). Condition (ii) also assumes there are a finite number of moment conditions. Condition (iii) assumes the feasible region is nonempty, and condition (v) assumes that the data is i.i.d. Finally, condition (iv) essentially requires that there exists a finite subset $\Theta' \subset \Theta$ such that the feasible regions for $\boldsymbol{\pi}$ described by the moment inequalities across values of $\theta \in \Theta$ is the same as the feasible regions for $\boldsymbol{\pi}$ described by the moment inequalities across values of $\theta \in \Theta'$. This is easily satisfied when Assumption 3.1 is imposed, as is described in the profiling section where Θ' is the collection of representative points.

The following result is provided in Gu and Russell (2021), and is stated here only for completeness. The result shows that the sample analog constraints and objective function in Theorem 2.2, combined with a slight (and vanishing) relaxation of the feasible region, leads to a consistent estimate of the identified set for counterfactual functionals, where consistency is measured with respect to the Hausdorff metric.[54]

Theorem E.1. *Suppose that Assumption E.1 holds. Then* $d_H(\Psi^*(\mathbb{P}_n, b_n), \Psi^*(P, 0)) = o_P(1)$*, where* b_n *is any positive sequence satisfying* $b_n^{-1} = o(\sqrt{n})$*.*

This result is connected to a number of results in the literature on partial identification going back to Manski and Tamer (2002). See Molchanov and Molinari (2018) Theorem 5.16 and the surrounding discussion. The required assumptions are easily checked for all models considered in the application section, where we use the sequence $b_n = c/\sqrt{\log(n)}$. All choices of the value $c > 0$ are theoretically valid. In the application section we set $c = 10^{-5}$, and found the results were not sensitive to this value.

[54] For two sets $A, B \subset \mathbb{R}^d$, the Hausdorff metric is defined as:

$$d_H(A, B) := \max\left\{\sup_{a\in A}\inf_{b\in B}||a-b||, \sup_{b\in B}\inf_{a\in A}||a-b||\right\}.$$