A Dual Approach to Wasserstein-Robust Counterfactuals

Jiaying Gu* *University of Toronto*

Thomas M. Russell† *Carleton University*

July 21, 2023

Abstract

We study the identification of scalar counterfactual parameters in partially identified structural models, paying particular attention to relaxing parametric distributional assumptions on the latent variables. We start by showing bounds on scalar counterfactual parameters can be constructed without parametric distributional assumptions by solving two infinite-dimensional optimization problems. Treating these as the primal problems, we use results from random set theory and convex analysis to reformulate the problems as finite-dimensional convex optimization problems involving the Aumann expectation of a random set, and then we derive the corresponding Fenchel dual problems. The dual problems can handle outcome variables and covariates with infinite support, and can easily allow a researcher to explore the sensitivity of their results to a baseline parametric distribution for the latent variables using the Wasserstein distance. We compare our approach to another dual approach by Christensen and Connault (2023), and propose an algorithm for estimation and inference. Finally, we apply the procedure to airline data from Ciliberto and Tamer (2009) and construct bounds on counterfactual market entry probabilities while exploring robustness to a parametric distribution for the latent variables.

Keywords: Counterfactuals, Partial Identification, Structural Models

We are grateful to seminar audiences at the CIREQ Econometrics Conference in Montreal, Conference on Models and Econometrics of Strategic Interactions at Vanderbilt University, Cornell University, Optimization-Conscious Econometrics Conference at the University of Chicago, University of British Columbia, University of Montreal, and the Workshop on Using Data to Make Decisions at Brown University. The 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, M5S 3G7, Canada. Email: jiaying.gu@utoronto.ca.

†Thomas M. Russell, Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada. Email: thomas.russell3@carleton.ca.

1 Introduction

Understanding how agents respond to hypothetical policy changes is important for both designing and implementing optimal policies. In complex environments with observational data, structural models can be used to simulate the effects of a counterfactual policy change and aid in the policy decision-making process. However, even simple structural models rely on strong assumptions on the latent variables and can be computationally challenging to estimate. Furthermore, when stringent modelling assumptions are relaxed or removed the model parameters may become partially identified, which typically exacerbates the computational challenges. A recent and growing literature has attempted to both relax assumptions on latent variables, and to address the resulting computational challenges associated with learning about counterfactual parameters.

This paper studies the problem of constructing bounds on scalar counterfactual parameters in partially identified structural models while relaxing parametric distributional assumptions on the latent variables. The initial motivation for the paper is the recent work of Gu et al. (2022) and Tebaldi et al. (2023), who show how to bound counterfactual parameters in structural models with finitely supported outcomes and covariates without making parametric distributional assumptions on the latent variables. These papers use the finite support of the observed random variables to effectively discretize the latent variables without any loss of identifying information. They then bound counterfactual parameters by solving a sequence of finite-dimensional linear programs. The current paper begins by showing that, using basic concepts from random set theory, the optimization problems in Gu et al. (2022) and Tebaldi et al. (2023) can be reformulated as abstract but finite-dimensional convex optimization problems over the Aumann expectation of a random set, even when the support of the observed random variables is infinite. We then show the associated dual problem can be written as the average of a sequence of simple optimization problems, often with only a few parameters and constraints. The dual approach avoids certain expensive pre-processing steps needed in Gu et al. (2022) and Tebaldi et al. (2023), making it more attractive when the observed random variables have large support. We also show how the dual problem can be used to investigate sensitivity to a baseline parametric distributional assumption on the latent variables using the Wasserstein distance.

This paper is inspired by the work of Christensen and Connault (2023), who also study the problem of bounding counterfactual quantities while relaxing assumptions on the latent variable distribution, and who also use an approach based on convex duality. Christensen and Connault (2023) consider robustness to parametric distributional assumptions on the latent variables using nonparametric neighborhoods of a baseline distribution defined using a $\phi-$divergence. Using infinite-dimensional convex duality, they show how to bound counterfactual functionals while allowing the latent variable distribution to vary in the neighborhood of the baseline distribution by

solving finite-dimensional optimization problems. In contrast, our dual approach is different, and relies on combining random set theory with finite-dimensional convex duality. We show the approach has advantages in models that are incomplete, or models that have variables with infinite support. We also introduce a notion of robustness using a Wasserstein neighborhood of a baseline distribution which has a number of comparative advantages relative to $\phi-$divergences.

The dual formulation presents its own challenges for estimation and inference, which we investigate in detail and propose a general-purpose algorithm. The optimization problems that characterize our bounds have three layers: an inner layer that profiles over latent and counterfactual variables, a convex middle layer that penalizes global constraint violations, and an outer layer that profiles over structural parameters. We show the inner layer can often be solved as a sequence of low dimensional linear programs (LPs) or mixed integer linear programs (MILPs). We show the value function of the inner problem is sub-differentiable under weak conditions, and propose a variant of stochastic subgradient descent to solve the middle problem. We then recommend Bayesian optimization to solve the outer problem profiling over structural parameters. Bayesian optimization was popularized by Jones et al. (1998) and Jones (2001), and advocated in the econometrics literature by Kaido et al. (2019).

After discussing the identification results and computational issues, we prove consistency of a simple plug-in estimator, and propose a new inference procedure tailored to our problem, requiring minimal assumptions and computational effort. In particular, we propose a confidence set for our counterfactual bounds that is computed by inverting a simple hypothesis test. The testing procedure is adapted from a recent specification testing procedure proposed by Marcoux et al. (2023), and is designed to recycle certain expensive-to-compute components from the estimation of our bounds.

Finally, we revisit the airline entry game of Ciliberto and Tamer (2009) and construct bounds on counterfactual market entry probabilities while exploring robustness to a parametric distribution for the latent variables. We show how the bounds change under different assumptions on the number of strategically interacting agents, as well as under different independence assumptions. The application also allows us to illustrate the practical performance of our proposed estimation algorithm.

1.1 Relevant Literature

A recent literature has focused on bounding counterfactual parameters in models that are partially identified (see Manski (2007), Chesher and Rosen (2021), Kalouptsidi et al. (2021), Gu and Russell (2022), Gu et al. (2022), Christensen and Connault (2023), Tebaldi et al. (2023)). The initial motivation for the current paper is the work of Gu et al. (2022) and Tebaldi et al. (2023), who show how to bound counterfactual parameters in a class of models with discrete outcomes and

covariates without making parametric distributional assumptions on the latent variables. However, the approach becomes infeasible when the observed random variables have infinite support. Under some assumptions the optimization problems considered in this paper can be seen as the Fenchel dual problems of the primal problems considered in Gu et al. (2022) and Tebaldi et al. (2023), and we show the dual is finite-dimensional even when the observed random variables have infinite support.

While Gu et al. (2022) and Tebaldi et al. (2023) allow for some semiparametric restrictions on the latent variable distribution, researchers may also be interested in exploring sensitivity of their results to a baseline parametric distribution for the latent variables. This connects the current paper to the work of Christensen and Connault (2023). Both this paper and the paper of Christensen and Connault (2023) are related to the literature on distributional robustness in optimization. Robustness can be achieved in a stochastic optimization problem by solving the problem for the worst case data-generating distribution, where the worst case is computed over a neighborhood of some baseline distribution. Our form of distributional robustness from a baseline distribution of latent variables can be interpreted using the Wasserstein distance, which has recently been employed in the econometrics literature by Chen et al. (2021), Adjaho and Christensen (2022), Fan et al. (2023), among others. The Wasserstein distance has also previously been considered in the context of robust optimization in the operations research and machine learning literature by Pflug and Wozabal (2007), Wozabal (2012), Shafieezadeh Abadeh et al. (2015), Mohajerin Esfahani and Kuhn (2018), Lee and Raginsky (2018), Sinha et al. (2018), Blanchet et al. (2019), and Blanchet and Murthy (2019). Our problem is conceptually different from the canonical distributional robustness setting: rather than considering distributional robustness with respect to the (joint) empirical distribution in prediction problems, we consider robustness with respect to a (marginal) baseline latent variable distribution to bound (and construct confidence sets for) counterfactual parameters in structural models. The resulting differences means many insights from the previous literature cannot be recycled for our setting.

Throughout the explanations and proofs, we make extensive use of results from random set theory and convex analysis. The main theoretical techniques in the paper are most similar to Beresteanu et al. (2011), who use a support function characterization of the expectation of a random set to provide a representation of the identified set of structural parameters for a general class of models. Ekeland et al. (2010), Henry and Mourifié (2013), Mohajerin Esfahani and Kuhn (2018) and Blanchet and Murthy (2019) obtain a similar dual in different contexts using Monge-Kantorovich duality from optimal transport theory. Related results are also obtained by Li (2021), who provides a characterization of the identified set for both structural and counterfactual parameters without requiring parametric distributional assumptions on the latent variables. Finally, Lee (2022) uses infinite-dimensional linear programming duality to derive bounds on certain functions of the random

coefficients in a linear dynamic random coefficient model. In contrast to all these papers, our focus is on bounding scalar counterfactual parameters in structural models without first recovering the identified set of structural parameters. The shift in focus allows us to formulate our bounding problem in terms of two optimization problems, which in turn motivate our assumptions and examples, and all of our results on estimation, inference, and computation.

2 Methodology

2.1 Main Assumptions and Examples

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space, and consider a setting where the researcher observes the random vectors $\boldsymbol{Y} : \Omega \to \mathcal{Y} \subset \mathbb{R}^{d_y}$ and $\boldsymbol{Z} : \Omega \to \mathcal{Z} \subset \mathbb{R}^{d_z}$, where both $\mathcal{Y}$ and $\mathcal{Z}$ are countable, and where $\boldsymbol{Y}$ represents a vector of endogenous outcome variables and $\boldsymbol{Z}$ represents a vector of covariates. The researcher's environment also includes a vector of latent variables $\boldsymbol{U} : \Omega \to \mathcal{U} \subset \mathbb{R}^{d_u}$, which are related to the observed random variables through a support restriction $\boldsymbol{U} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta_0)$ a.s. Here $\theta_0 \in \Theta \subset \mathbb{R}^{d_\theta}$ represents a vector of true but unknown structural parameters, and $\mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is a set representing the values of $\boldsymbol{U} \in \mathcal{U}$ that are possible given the researcher's model, given the structural parameters are fixed at θ, and given the observed vector is $(\boldsymbol{Y}, \boldsymbol{Z})$. Alternatively, the support restrictions on $\boldsymbol{U}$ can be implicitly defined by requiring $\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{Z}, \boldsymbol{U}, \theta_0)$ a.s., where $\mathcal{Y}(\boldsymbol{Z}, \boldsymbol{U}, \theta)$ represents the set of model-predicted outcomes.[1] For now we do not impose any parametric distributional assumptions on the latent vector $\boldsymbol{U}$, although we allow the researcher to constrain the distribution of $\boldsymbol{U}$ using a vector of moment functions. Later we also show how to explore sensitivity to a baseline parametric distribution using the Wasserstein distance. This discussion, as well as some other technical conditions, are formalized in the following assumption.

Assumption 2.1. *There exists a complete and non-atomic probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and random vectors $\boldsymbol{Y} : \Omega \to \mathcal{Y}$, $\boldsymbol{Z} : \Omega \to \mathcal{Z}$, and $\boldsymbol{U} : \Omega \to \mathcal{U}$, where $\mathcal{Y} \subset \mathbb{R}^{d_y}$ and $\mathcal{Z} \subset \mathbb{R}^{d_z}$ are countable, and $\mathcal{U} \subset \mathbb{R}^{d_u}$, satisfying $\boldsymbol{U} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta_0)$ almost surely for some $\theta_0 \in \Theta \subset \mathbb{R}^{d_\theta}$, where $\mathcal{U}(\,\cdot\,, \theta) : \mathcal{Y} \times \mathcal{Z} \to 2^{\mathcal{U}}$ is a nonempty multifunction such that $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{y}, \boldsymbol{z}, \theta)\}$ is measurable for each $\theta \in \Theta$. Furthermore, $\mathbb{E}[\boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \theta_0)] = \boldsymbol{0}$ for a measurable vector-valued function $\boldsymbol{m} : \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \Theta \to \mathbb{R}^{d_m}$ with $\boldsymbol{u} \mapsto \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta)$ a continuous map for each $(\boldsymbol{y}, \boldsymbol{z}, \theta) \in \mathcal{Y} \times \mathcal{Z} \times \Theta$.*

Remark 2.1. *Restricting the probability space to be non-atomic is a technical assumption, and does not restrict the observed random variables, or any other random variables defined on $(\Omega, \mathfrak{A}, \mathbb{P})$, to be continuous.*

[1] The characterizations are dual to one another in following sense:

$$\boldsymbol{U} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta) \text{ a.s.} \iff \boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{Z}, \boldsymbol{U}, \theta) \text{ a.s.}$$

Remark 2.2. *Assumption 2.1 permits restrictions on the moments of a latent variable U_i, such as $\mathbb{E}[U_i] = 0$ and $\mathbb{E}[U_i^2] = 1$, as well as restrictions on the dependence between the latent variables $\boldsymbol{U}_i$ and the covariates, such as $\mathbb{E}[U_i Z_i] = 0$ combined with $\mathbb{E}[U_i] = 0$.*

Remark 2.3. *Restricting $\mathcal{Y} \times \mathcal{Z}$ to be countable is used to handle certain measurability issues without needing to introduce high-level assumptions on the model. While a theoretical restriction, all finite-precision data has countable support. The assumption is used only in the proof of Lemma S.2.2 in the Supplemental Material, and any other assumptions sufficient for Lemma S.2.2 are valid substitutions.*

We assume the researcher's main objective is to learn about the expected value of a scalar function $\varphi(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)$, where $\boldsymbol{Y}^\star : \Omega \to \mathcal{Y} \subset \mathbb{R}^{d_y}$ is a counterfactual outcome vector that results after an intervention in the environment. The random vector $\boldsymbol{Y}^\star$ is constrained to satisfy $\boldsymbol{Y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{U}, \theta_0)$ a.s. In particular, we assume the researcher's counterfactual of interest can be imposed as restrictions on the possible values the counterfactual outcome variable can take, where the restrictions can depend on the observed covariates, the latent variables, and the values of the structural parameters. Throughout, we refer to $\mathbb{E}[\varphi(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)]$ as the *counterfactual functional*, and most of the effort in the paper will be spent on constructing a tractable procedure to bound this parameter. We formalize the restrictions on the counterfactual environment in the following assumption.

Assumption 2.2. *There exists a random vector $\boldsymbol{Y}^\star : \Omega \to \mathcal{Y}$ satisfying $\boldsymbol{Y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{U}, \theta_0)$ almost surely for the same $\theta_0 \in \Theta$ from Assumption 2.1, where $\mathcal{Y}^\star(\,\cdot\,, \theta) : \mathcal{Z} \times \mathcal{U} \to 2^{\mathcal{Y}}$ is an nonempty multifunction such that $(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}) \mapsto \mathbb{1}\{\boldsymbol{y}^\star \in \mathcal{Y}(\boldsymbol{z}, \boldsymbol{u}, \theta)\}$ is measurable for each $\theta \in \Theta$. Furthermore, the researcher's objective function $\varphi(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)$ is measurable for each $\theta \in \Theta$ and the map $(\boldsymbol{y}^\star, \boldsymbol{u}) \mapsto \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta)$ is continuous for each $(\boldsymbol{z}, \theta)$.*

To fix ideas, we now provide three examples of different models and counterfactuals that fit into the framework.

Example 1 (Multinomial Choice)**.** *Consider a model of multinomial choice, where consumers select among J alternatives. Suppose the utility obtained from choice $j \in \mathcal{J} = \{1, \ldots, J\}$ is:*

$$\pi_j(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \boldsymbol{z}_{ij}^\top \theta_j + u_{ij},$$

where $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ})$ is a vector of individual-choice specific attributes, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iJ})$ is a vector of latent utility shocks, and $\theta = (\theta_1, \ldots, \theta_J)$ is a vector of structural parameters common to all consumers. Let $Y_i \in \mathcal{J}$ denote the choice of consumer i, and suppose each consumer selects the

alternative that obtains the highest utility. The $\boldsymbol{Y}-$level set can be written as:

$$\mathcal{Y}(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \left\{ y \in \mathcal{J} : y = \arg\max_{j \in \mathcal{J}} \pi_j(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \right\},$$

and the $\boldsymbol{U}$-level set written as:

$$\mathcal{U}(y_i, \boldsymbol{z}_i, \theta) = \left\{ \boldsymbol{u} \in \mathbb{R}^J : \boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{iy_i} \geq \boldsymbol{z}_{iy'}^\top \theta_j + u_{iy'}, \forall y' \neq y_i \right\}.$$

Now consider a counterfactual where the vector $\boldsymbol{z}_i$ is replaced by the vector $\check{\boldsymbol{z}}_i = \check{\boldsymbol{z}}_i(\boldsymbol{z}_i)$ (a known function of $\boldsymbol{z}_i$), and suppose we are interested in the resulting change in social surplus. The counterfactual $\boldsymbol{Y}-$level set is given by:

$$\mathcal{Y}^\star(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \left\{ y^\star \in \mathcal{J} : y^\star = \arg\max_{j \in \mathcal{J}} \pi_j(\check{\boldsymbol{z}}_i, \boldsymbol{u}_i, \theta) \right\},$$

and our counterfactual parameter of interest is:

$$\mathbb{E}[\varphi(\boldsymbol{y}_i^\star, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta)] = \mathbb{E}\left[\max_{j \in \mathcal{J}} \{\check{\boldsymbol{z}}_{ij}^\top \theta_j + u_{ij}\} - \max_{j \in \mathcal{J}} \{\boldsymbol{z}_{ij}^\top \theta_j + u_{ij}\} \right].$$

Example 2 (Binary Game)**.** *Consider a binary game with $K = 2$ players with complete information and pure strategy Nash equilibria. Each player chooses a binary action, and receives a payoff which depends on the actions of the other player. An outcome of the i^{th} game is a vector $\boldsymbol{y}_i \in \{0,1\}^2$ specifying the actions chosen by each player. The payoffs received by each player in the i^{th} game are given by:*

$$\pi_1(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i; \theta) = y_{i1}(\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 - u_{i1}), \qquad \pi_2(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i; \theta) = y_{i2}(\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 - u_{i2}),$$

where $\theta := (\beta_1, \beta_2, \delta_1, \delta_2)$ is a vector of fixed structural parameters, $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ are player-specific observable covariates, and $\boldsymbol{u}_i = (u_{i1}, u_{i2})$ represents payoff-relevant unobserved heterogeneity. Under pure strategy Nash equilibria the observed vector of entry decisions must belong to the set:

$$\mathcal{Y}(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \left\{ \boldsymbol{y}_i \in \{0,1\}^2 : \begin{array}{l} y_{i1} = \mathbb{1}\{\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 \geq u_{i1}\} \\ y_{i2} = \mathbb{1}\{\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 \geq u_{i2}\} \end{array} \right\},$$

and thus the vector of latent variables $\boldsymbol{u}_i$ must belong to the set:

$$\mathcal{U}(\boldsymbol{y}_i, \boldsymbol{z}_i, \theta) = \left\{ \boldsymbol{u}_i \in \mathbb{R}^2 : \begin{array}{l} y_{i1} = \mathbb{1}\{\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 \geq u_{i1}\} \\ y_{i2} = \mathbb{1}\{\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 \geq u_{i2}\} \end{array} \right\}.$$

Now consider a counterfactual that replaces $(\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ with $(\check{\boldsymbol{z}}_{i1}, \check{\boldsymbol{z}}_{i2})$ and forces player 2 to play

action 0. *For this counterfactual the counterfactual actions* $\boldsymbol{y}_i^\star$ *belong to the set:*

$$\mathcal{Y}^\star(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \left\{ \boldsymbol{y}_i^\star \in \{0,1\}^2 : \begin{array}{l} y_{i1}^\star = \mathbb{1}\{\check{\boldsymbol{z}}_{i1}^\top \beta_1 + y_{i2}^\star \delta_1 \geq u_{i1}\} \\ y_{i2}^\star = 0 \end{array} \right\}.$$

A researcher interested in the counterfactual entry probability of player 1 can set $\varphi(\boldsymbol{y}_i^\star, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = y_{i1}^\star$*, and can use our method to bound* $\mathbb{E}[\varphi(\boldsymbol{Y}_i^\star, \boldsymbol{Z}_i, \boldsymbol{U}_i, \theta)]$.

Example 3 (Network Formation)**.** *Consider observations on* n *pairwise stable networks, formed with either transferable or non-transferable utility.*[2] *Suppose* $K \geq 3$ *individuals consider forming a network* $\mathcal{N}_i \subset [K] \times [K]$*, where* $(k, k') \in \mathcal{N}_i$ *if there exists a link between individuals* k *and* k' *in network* $\mathcal{N}_i$*. The network* $\mathcal{N}_i$ *can also be represented by a vector* $\boldsymbol{y}_i \in \{0,1\}^{K(K-1)/2}$ *where* $\boldsymbol{y}_i = (y_{i12}, y_{i13}, \ldots, y_{i(n-1)n})^\top$*, and* $y_{ikk'} = 1$ *if and only if there exists a link between individuals* k *and* k' *in network* i*. Links are formed between individuals* k *and* k' *if the marginal utility of doing so is positive for both individuals. Assume the marginal utility to agent* k *from forming a link with individual* k' *is given by:*

$$\Delta\pi_{kk'}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \underbrace{\theta_1 + \boldsymbol{z}_{ik}^\top \theta_2 + ||\boldsymbol{z}_{ik} - \boldsymbol{z}_{ik'}||\theta_3 + u_{ikk'}}_{\textit{Direct Utility}} + \underbrace{\frac{\theta_4}{K-2} \sum_{j \neq k,k'} y_{ijk'}}_{\textit{Indirect Connections}} + \underbrace{\frac{\theta_5}{K-2} \sum_{j \neq k,k'} y_{ijk} y_{ijk'}}_{\textit{Mutual Connections}}.$$

Set $\mathcal{Y} = \{0,1\}^{K(K-1)/2}$*. Then* $\boldsymbol{y}_i$ *arises from a pairwise stable undirected network under non-transferable utility if and only if it belongs to the set:*

$$\mathcal{Y}(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) := \left\{ \boldsymbol{y} \in \mathcal{Y} : \boldsymbol{y} = \begin{bmatrix} \mathbb{1}\{\Delta\pi_{12}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \mathbb{1}\{\Delta\pi_{21}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \\ \mathbb{1}\{\Delta\pi_{13}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \mathbb{1}\{\Delta\pi_{31}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \\ \vdots \\ \mathbb{1}\{\Delta\pi_{(K-1)K}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \mathbb{1}\{\Delta\pi_{K(K-1)}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \end{bmatrix} \right\}.$$

In contrast, $\boldsymbol{y}_i$ *arises from a pairwise stable undirected network under transferable utility if and only if it belongs to the set:*

$$\mathcal{Y}(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) := \left\{ \boldsymbol{y} \in \mathcal{Y} : \boldsymbol{y} = \begin{bmatrix} \mathbb{1}\{\Delta\pi_{12}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) + \Delta\pi_{21}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \\ \mathbb{1}\{\Delta\pi_{13}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) + \Delta\pi_{31}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \\ \vdots \\ \mathbb{1}\{\Delta\pi_{(K-1)K}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) + \Delta\pi_{K(K-1)}(\boldsymbol{y}, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\} \end{bmatrix} \right\}.$$

The set $\mathcal{U}(\boldsymbol{y}_i, \boldsymbol{z}_i, \theta)$ *is defined analogously. A similar approach can be used to construct the* $\boldsymbol{Y}-$*level and* $\boldsymbol{U}-$*level sets for directed networks (e.g. Gualdani (2021)), models of dyadic link formation (e.g. Gao et al. (2022)), and matching models (e.g. Gualdani and Sinha (2020)). Now suppose*

[2] De Paula (2020) provides a detailed overview of recent advances in econometric models of network formation.

that $\boldsymbol{z}_i$ are replaced with counterfactual covariate values $\check{\boldsymbol{z}}_i = \check{\boldsymbol{z}}_i(\boldsymbol{z}_i)$, and suppose the researcher wishes to study the effect of the intervention on the average number of links between agents of type $\mathcal{T}_1(\boldsymbol{z}_i) \subset [K]$ with agents of type $\mathcal{T}_2(\boldsymbol{z}_i) \subset [K]$. The researcher will set:

$$\varphi(\boldsymbol{y}_i^\star, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \sum_{k \in \mathcal{T}_1(\boldsymbol{z}_i)} \sum_{k' \in \mathcal{T}_2(\boldsymbol{z}_i)} y_{ikk'}^\star,$$

and can then use our method to bound the value of $\mathbb{E}[\varphi(\boldsymbol{Y}_i^\star, \boldsymbol{Z}_i, \boldsymbol{U}_i, \theta)]$.

Remark 2.4. *There can be multiple equivalent ways to express the counterfactual of interest using the framework, some of which may or may not satisfy Assumption 2.2. For instance, if $Y_i^\star = \mathbb{1}\{U_i \geq 0\}$, this can be expressed by setting $\varphi(y_i^\star, u_i, \theta) = \mathbb{1}\{u_i \geq 0\}$, omitting the set $\mathcal{Y}^\star(u_i, \theta)$ altogether. However, this does not satisfy Assumption 2.2, since $\varphi(y_i^\star, u_i, \theta)$ is not continuous in u_i. Alternatively, setting $\varphi(y_i^\star, u_i, \theta) = y_i^\star$ and $\mathcal{Y}^\star(u_i, \theta) = \{y_i^\star \in \{0,1\} : y_i^\star = \mathbb{1}\{u_i \geq 0\}\}$ ensures Assumption 2.2 is satisfied.*

2.2 Identification Results

Our target of interest is the identified set for the counterfactual functional, which is defined next.

Definition 2.1. *Under Assumptions 2.1 and 2.2, the identified set Φ^* for the functional $\varphi : \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \Theta \to \mathbb{R}$ is the set of all values $\overline{\varphi} \in \mathbb{R}$ for which there exists random vectors $\widetilde{\boldsymbol{U}} : \Omega \to \mathcal{U}$ and $\widetilde{\boldsymbol{Y}}^\star : \Omega \to \mathcal{Y}$ satisfying the following conditions for some $\theta \in \Theta$: (i) $\widetilde{\boldsymbol{U}} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ almost surely, (ii) $\widetilde{\boldsymbol{Y}}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \widetilde{\boldsymbol{U}}, \theta)$ almost surely, (iii) $\mathbb{E}[\boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \widetilde{\boldsymbol{U}}, \theta)] = \boldsymbol{0}$, and (iv) $\bar{\varphi} = \mathbb{E}[\varphi(\widetilde{\boldsymbol{Y}}^\star, \boldsymbol{Z}, \widetilde{\boldsymbol{U}}, \theta)]$. Furthermore, the values of $\theta \in \Theta$ for which these conditions are satisfied for some $\overline{\varphi} \in \mathbb{R}$ is called the identified set of structural parameters, and is denoted as Θ^*.*

The conditions in Definition 2.1 ensure the identified set Φ^* consists of all and only those values of the counterfactual parameter $\bar{\varphi} \in \mathbb{R}$ that can be rationalized by a vector of latent variables $\widetilde{\boldsymbol{U}}$ satisfying the model restrictions from Assumption 2.1 and a vector of counterfactual outcome variables $\widetilde{\boldsymbol{Y}}^\star$ satisfying the counterfactual support restrictions from Assumption 2.2 for some $\theta \in \Theta^*$. Note that without additional restrictions the identified set Φ^* can be open and disconnected. Thus, throughout the paper we focus on constructing the closed convex hull of Φ^*, which is an interval with endpoints characterized by two optimization problems.

To appreciate the theoretical challenges associated with bounding a counterfactual functional under our assumptions, first define the functions:

$$\delta_1(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta) := \mathbb{1}\{\boldsymbol{u} \notin \mathcal{U}(\boldsymbol{y}, \boldsymbol{z}, \theta)\}, \qquad \delta_2(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) := \mathbb{1}\{\boldsymbol{y}^\star \notin \mathcal{Y}^\star(\boldsymbol{z}, \boldsymbol{u}, \theta)\}. \tag{2.1}$$

Then the lower bound on the closed convex hull of Φ^* is given by the value of the following

optimization problem:

$$\inf_{\theta\in\Theta^*}\inf_{(\boldsymbol{U},\boldsymbol{Y}^\star)\in\mathcal{M}_u\times\mathcal{M}_y}\mathbb{E}[\varphi(\boldsymbol{Y}^\star,\boldsymbol{Z},\boldsymbol{U},\theta)]\text{ s.t. }\mathbb{E}[\delta_1(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\theta)]=0,\ \mathbb{E}[\delta_2(\boldsymbol{Y}^\star,\boldsymbol{Z},\boldsymbol{U},\theta)]=0,\ \mathbb{E}[\boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\theta)]=\mathbf{0}. \tag{2.2}$$

where $\mathcal{M}_u$ and $\mathcal{M}_y$ denote the set of all random vectors (measurable functions) from $(\Omega,\mathfrak{F})$ to $(\mathcal{U},\mathfrak{B}(\mathcal{U}))$ and $(\mathcal{Y},\mathfrak{B}(\mathcal{Y}))$. In particular, given the random vector $(\boldsymbol{Y},\boldsymbol{Z})$ and a value of the structural parameters $\theta\in\Theta^*$ the inner optimization problem searches over all possible latent variable vectors $\boldsymbol{U}$ and counterfactual outcome vectors $\boldsymbol{Y}^\star$ for the pair that minimizes the counterfactual functional, subject to the support constraints $\boldsymbol{U}\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ and $\boldsymbol{Y}^\star\in\mathcal{Y}(\boldsymbol{Z},\boldsymbol{U},\theta)$ a.s., and the additional moment conditions. Since the upper bound on the closed convex hull of Φ^* has a similar formulation, we focus on the lower bound throughout.

The obvious challenge to solving (2.2) is that it requires an infeasible search over all possible random vectors. In recent work, Gu et al. (2022) and Tebaldi et al. (2023) show that in a certain class of models with finitely-supported outcomes and covariates the latent variables can be discretized without any loss of identifying information, making the problem (2.2) finite-dimensional. However, they require the model to have a particular index structure, and it is difficult to impose constraints on the latent variable distribution in their framework using moment conditions. Outside of their class of models, or if either $\boldsymbol{Y}$ or $\boldsymbol{Z}$ has infinite support, the optimization problems in (2.2) generally remain intractable.

Since the problem (2.2) is the first and most natural way to formulate the lower bound on the closed convex hull of Φ^*, we refer to it as the *primal problem*. To make progress, we first rewrite the primal problem in terms of an abstract finite-dimensional convex optimization problem involving the Aumann expectation of a random set, and then we derive the corresponding Fenchel dual problem and show that it is more tractable in many cases.

Step 1: Rewriting the Primal

Our analysis relies on representing the model through the following set:

$$\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta):=\text{cl}\left\{(x_1,\boldsymbol{x}_2,x_3,x_4)\in\mathbb{R}^{d_m+3}:\begin{array}{l}x_1=\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta),\\ \boldsymbol{x}_2=\boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta),\\ x_3=\delta_1(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta),\\ x_4=\delta_2(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta),\end{array}\quad(\boldsymbol{u},\boldsymbol{y}^\star)\in\mathcal{U}\times\mathcal{Y}\right\}, \tag{2.3}$$

where $\delta_1(\boldsymbol{y},\boldsymbol{z},\boldsymbol{u},\theta)$ and $\delta_2(\boldsymbol{y},\boldsymbol{z},\boldsymbol{u},\theta)$ are as defined in (2.1). Fixing $\theta\in\Theta$, given a realization of $(\boldsymbol{Y}(\omega),\boldsymbol{Z}(\omega))$ the set (2.3) traces out all possible values of the counterfactual functional, the moment functions, and the functions δ_1 and δ_2 that are possible *for some* pair $(\boldsymbol{u},\boldsymbol{y}^\star)\in\mathcal{U}\times\mathcal{Y}$. It

represents *all* the support restrictions implied by the researcher's model that matter for solving the primal problem. We require the following regularity conditions for our main identification results.

Assumption 2.3. *For every* $\theta \in \Theta$, $\mathbb{E}\left[\sup\{||\boldsymbol{X}|| : \boldsymbol{X} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)\}\right] < \infty$.

Lemma 2.1. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then for every* $\theta \in \Theta$, $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ *is weakly measurable and almost surely nonempty, and thus is a random closed set.*

The definition of a weakly measurable multifunction and of a random closed set are provided in Section S.1 of the Supplemental Material. Lemma 2.1 demonstrates that, under Assumptions 2.1, 2.2, and 2.3, the multifunction $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ satisfies the measurability requirements to be designated as a random closed set. Assumption 2.3 is similar to Assumption 2.3 in Beresteanu et al. (2011), and imposes a uniform integrability requirement on the selections of the random set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Random sets that satisfy this condition are called *integrably bounded*. Importantly, the condition implies $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is a random compact set, which may be restrictive in some contexts.[3]

The *selections* of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ are random vectors $\boldsymbol{X} : \Omega \to \mathbb{R}^{d_m+3}$ with realizations satisfying $\boldsymbol{X}(\omega) \in \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$ a.s., and the set of all selections of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is denoted by $\text{Sel}(\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))$. With this concept in hand the primal problem admits a convenient reformulation.

Proposition 2.1. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then the problems:*

$$\inf_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathbb{E}[\varphi(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)] \text{ s.t. } \mathbb{E}[\delta_1(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = 0,\ \mathbb{E}[\delta_2(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = 0, \quad (2.4)$$
$$\mathbb{E}[\boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = \boldsymbol{0},$$

and:

$$\inf_{(X_1, \boldsymbol{X}_2, X_3, X_4) \in Sel\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \mathbb{E}[X_1] \text{ s.t. } \mathbb{E}[\boldsymbol{X}_2] = \boldsymbol{0},\ \mathbb{E}[X_3] = 0,\ \mathbb{E}[X_4] = 0, \quad (2.5)$$

have the same value for all $\theta \in \Theta^*$.

Proposition 2.1 shows the optimization problem in (2.4) can be rewritten as an optimization problem over random variables $(X_1, \boldsymbol{X}_2, X_3, X_4)$ that can be realized as measurable selections from the random set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Intuitively, the objective function, constraint functions, and random vectors $(\boldsymbol{U}, \boldsymbol{Y}^\star)$ from (2.4) have been absorbed into the random set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Absorbing $(\boldsymbol{U}, \boldsymbol{Y}^\star)$ is possible by the construction of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$, since it represents all values of the objective function and constraint functions that are possible *for some* values of the random vectors $(\boldsymbol{U}, \boldsymbol{Y}^\star)$. Formally the proof of Proposition 2.1 shows a random vector $\boldsymbol{X}$ is a selection of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ satisfying the

[3]See Molchanov (2017) p. 227. For instance, the assumption can fail if both the set $\mathcal{U}(\boldsymbol{y}, \boldsymbol{z}, \theta)$ is unbounded, and the moment functions $\boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta)$ are unbounded in $\boldsymbol{u}$ for some $\theta \in \Theta$ and some $(\boldsymbol{y}, \boldsymbol{z})$ occurring with positive probability. In these cases our theoretical results can be applied after an appropriate truncation of the moment conditions or the support of the latent variables.

constraints and obtaining a certain value in program (2.5) if and only if there exists $(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y$ satisfying the constraints and obtaining the same value in program (2.4).

Written using selections and their expectations, the program (2.5) can be further simplified using concepts from random set theory. In particular, we require the concept of the *Aumann expectation* of a random set:

$$\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta) := \{\mathbb{E}\boldsymbol{X} : \boldsymbol{X} \in \text{Sel}(\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))\}.$$

In words, the Aumann expectation of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is a set that collects the expected values of all selections of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Since the program (2.5) is written using expectations of selections from $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$, it can be written using the Aumann expectation.

Proposition 2.2. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then the program:*

$$\inf_{(a_1, \mathbf{0}, 0, 0) \in \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} a_1, \tag{2.6}$$

has the same value as the program (2.5) for all $\theta \in \Theta^*$.

The program (2.6) searches over all selections of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ for the one whose first element (corresponding to the counterfactual functional) has the smallest expected value, while constraining all other elements (corresponding to the constraints) to have zero expected value. The program is especially simple since the Aumann expectation of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ has a number of desirable properties under our maintained assumptions.

Lemma 2.2. *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Then the Aumann expectation is nonempty, closed, convex, and bounded.*

Convexity of $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ implies the problem in (2.6) is a finite-dimensional convex optimization problem, which searches over vectors $\boldsymbol{a} \in \mathbb{R}^{d_m+3}$ for the minimum common point between the set $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ and the axis corresponding to the first element a_1. An illustration is provided in Figure 1. In practice, it is difficult to explicitly construct the Aumann expectation $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$, so (2.6) remains an abstract and intractable program. However, under some regularity conditions, this finite-dimensional convex optimization problem has a tractable finite-dimensional dual problem.

Step 2: Deriving the Dual

Given the simple form of the convex optimization problem in (2.6), the corresponding dual problem has a straightforward geometric interpretation, which is illustrated in Figure 1.[4] As a closed and convex set, the Aumann expectation $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ can be characterized as the intersection

[4]See the discussion of the "min-common point" and "max-crossing" problems in Bertsekas (2009).

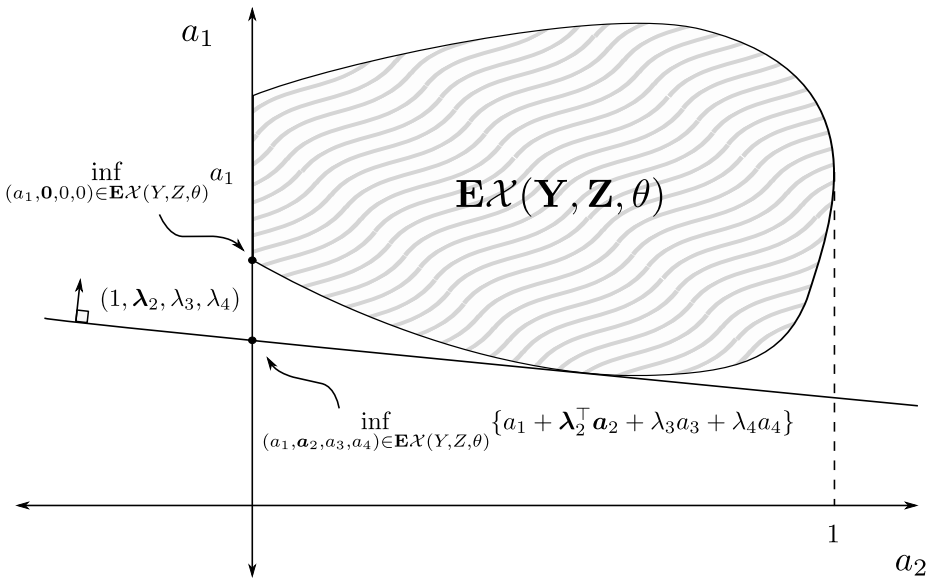


Figure 1: An illustration of the projection of $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ on the $(a_1,a_2)-$plane. The figure illustrates the minimum common point between the set $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ and the a_1-axis (the value of the primal problem (2.6)), as well as the maximum point at which a hyperplane with normal $(1,\boldsymbol{\lambda}_2,\lambda_3,\lambda_4)$ that contains $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ in its positive half-space can cross the a_1-axis (the dual problem (2.8)).

of all closed halfspaces corresponding to its supporting hyperplanes. Now consider the set of all non-vertical (with respect to the a_1-axis) hyperplanes H such that $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ is contained in the positive closed half-space defined by H.[5] After normalizing the coefficient on a_1 to 1, these hyperplanes are all of the form:

$$H=\{(a_1,\boldsymbol{a}_2,a_3,a_4): a_1+\boldsymbol{\lambda}_2^\top \boldsymbol{a}_2+\lambda_3 a_3+\lambda_4 a_4=b\},$$

where b denotes the point at which H intersects the a_1-axis. Then the set $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ is contained in the positive closed halfspace defined by H if and only if:

$$b\leq \inf_{\boldsymbol{a}\in\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \left\{a_1+\boldsymbol{\lambda}_2^\top \boldsymbol{a}_2+\lambda_3 a_3+\lambda_4 a_4\right\}. \tag{2.7}$$

The right side of (2.7) is the maximum point at which a hyperplane with normal vector $(1,\boldsymbol{\lambda}_2,\lambda_3,\lambda_4)$ containing $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ in its positive halfspace can cross the a_1-axis. Figure 1 provides an illustration. With this background, the dual problem to (2.6) is given by:

$$\sup_{(\boldsymbol{\lambda}_2,\lambda_3,\lambda_4)\in\mathbb{R}^{dm+2}} \inf_{\boldsymbol{a}\in\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \left\{a_1+\boldsymbol{\lambda}_2^\top \boldsymbol{a}_2+\lambda_3 a_3+\lambda_4 a_4\right\}. \tag{2.8}$$

[5]Given a hyperplane $H=\left\{\boldsymbol{x}\in\mathbb{R}^d:\langle \boldsymbol{x},\boldsymbol{a}\rangle=b\right\}$, the positive closed half-space defined by H is the set $H_+:=\left\{\boldsymbol{x}\in\mathbb{R}^d:\langle \boldsymbol{x},\boldsymbol{a}\rangle\geq b\right\}$.

In other words, the dual problem returns the largest value on the a_1-axis that can be crossed by a non-vertical hyperplane containing $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ in its positive halfspace.[6] As suggested by Figure 1, under some conditions the dual problem will have the same value as problem (2.6). Inspecting (2.8), we see that it can be rewritten as:

$$\sup_{\boldsymbol{q} \in \mathbb{R}^{d_m+3}} -s(-\boldsymbol{q}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)), \tag{2.9}$$

where $\boldsymbol{q} = (1, \boldsymbol{\lambda}_2, \lambda_3, \lambda_4)$ and $s(\,\cdot\,, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)) : \mathbb{R}^{d_m+3} \to \mathbb{R}$ is the support function of the convex set $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. After some derivation, the dual problem has an especially simple form.

Theorem 2.1. *Suppose Assumptions 2.1, 2.2, and 2.3 hold, fix any $\theta \in \Theta^*$ and consider the following primal problem:*

$$\inf_{(a_1, \mathbf{0}, 0, 0) \in \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} a_1. \tag{2.10}$$

Then (i) the infimum in (2.10) in attained and is finite, (ii) the corresponding Fenchel dual problem to (2.10) is:

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \int \inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \inf_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} \{\varphi(\boldsymbol{y}^\star, \boldsymbol{Z}, \boldsymbol{u}, \theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \theta) \rangle\} \, d\mathbb{P}, \tag{2.11}$$

where the outer supremum is attained, and (iii) the duality gap between (2.10) and (2.11) is zero.

The definition of the Fenchel dual is provided in Section S.1 of the Supplemental Material. The result shows the abstract problem (2.10) is equal in value to the average of a number of penalized "local problems" over $\boldsymbol{u}$ and $\boldsymbol{y}^\star$. The dual program can thus be viewed as a decentralized version of the primal. The local problems operate independently of each other, but the value of $\boldsymbol{\lambda}$ serves as a coordination device across the local problems ensuring that, in aggregate, the solutions to the local problems satisfy the model moment conditions. In the next section we show the local problems in our examples can be formulated as low-dimensional LPs or MILPs. Furthermore, the dual formulation does not require the outcomes $\boldsymbol{Y}$ or covariates $\boldsymbol{Z}$ to have finite support, providing a practical advantage over previous approaches to bounding counterfactual functionals.

Crucial to the proof of Theorem 2.1 is the interchange of an expectation and a supremum; in particular, we require the support function of the Aumann expectation of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is equal to the expected support function of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$:

$$-s(-\boldsymbol{q}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)) = -\mathbb{E}\left[\sup_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \langle -\boldsymbol{q}, \boldsymbol{x} \rangle\right], \tag{2.12}$$

for every $\boldsymbol{q} \in \mathbb{R}^{d_m+3}$. The interchange was also crucial in the development of support-function

[6] The hyperplane that obtains this value may not be unique, but this is not needed for strong duality.

estimators by Beresteanu et al. (2011), and holds for an integrably bounded random closed set when the underlying probability space is non-atomic.[7] Combining (2.9) and (2.12) the form of the dual program (2.11) follows almost immediately.

Remark 2.5. *Due to the particular form of the problem* (2.10)*, equality between* (2.10) *and* (2.11) *is guaranteed by closedness, boundedness, and convexity of* $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ *(by Lemma 2.2), and requires no separate constraint qualification (see Bertsekas (2009) Proposition 4.3.2).*

The previous result shows the form of the dual problem for a fixed $\theta \in \Theta^*$, but is silent on how to construct the identified set Θ^* for the structural parameters. In most realistic models, practically constructing the identified set Θ^* can be computationally demanding. The following result shows it is not neccessary to first construct the identified set Θ^* in order to construct the closed convex hull of the identified set Φ^* for the counterfactual functional.

Theorem 2.2. *Suppose Assumptions 2.1, 2.2, and 2.3 hold, and define:*

$$h(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{y}^\star, \boldsymbol{\lambda}, \theta) := \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta) \rangle.$$

Then $\overline{conv}(\Phi^*) = [\varphi_{\ell b}, \varphi_{ub}]$*, where:*

$$\varphi_{\ell b} := \inf_{\theta \in \Theta} \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \int \inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \inf_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} h(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{y}^\star, \boldsymbol{\lambda}, \theta) \, d\mathbb{P}, \tag{2.13}$$

$$\varphi_{ub} := \sup_{\theta \in \Theta} \inf_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \int \sup_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \sup_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} h(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{y}^\star, \boldsymbol{\lambda}, \theta) \, d\mathbb{P}. \tag{2.14}$$

Theorem 2.1 states that the closed convex hull of the identified set Φ^* from Definition 2.1 can be computed as the solution to two optimization problems. Importantly, Theorem 2.1 suggests there is no need to compute the full identified set Θ^* of structural parameters in order to bound certain counterfactual quantities. The result follows since the program in (2.11) (or the analogous upper bound program) will return $+\infty$ ($-\infty$) for any $\theta \notin \Theta^*$, but will return a finite value for each $\theta \in \Theta^*$, as shown in Theorem 2.1. In this way the outer optimization over $\theta \in \Theta^*$ in the programs (2.13) and (2.14) will avoid values of $\theta \notin \Theta$, since the outer optimization problems can always improve the objective function by considering $\theta \in \Theta^*$. Compared to a brute-force approach which evaluates the counterfactual at every $\theta \in \Theta^*$, the characterization suggested by Theorem 2.2 can dramatically simplify computation in practice, something we demonstrate in Section 5.

Remark 2.6. *The bounds* (2.13) *and* (2.14) *also simplify in some special cases. For instance, in complete models, or in models where* $\mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{U}, \theta)$ *is otherwise known to be single-valued for every* θ*, the inner optimization problems over* $\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)$ *can be removed in* (2.13) *and* (2.14)*.*

[7] See Molchanov (2017) Theorem 2.1.35. In particular, the result requires $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is closed, which is guaranteed when $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is integrably bounded.

Similarly, if $\mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is single-valued it can also be removed from the optimization problems in (2.13) and (2.14).

2.3 Exploring Robustness to Parametric Distributional Assumptions

A benefit of the dual formulation for counterfactual bounds is that it can easily allow researchers to explore sensitivity to parametric distributional assumptions. For this purpose, we require the following additional assumption.

Assumption 2.4. *For known $q \in [1, \infty)$, known $\rho \in \mathbb{R}_+$, and known distribution $P_{\boldsymbol{U}'}$ on $\mathcal{U}$, there exists a random vector $\boldsymbol{U}' : \Omega \to \mathcal{U} \subset \mathbb{R}^{d_u}$ with distribution $P_{\boldsymbol{U}'}$ such that $(\mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}'||_q^q])^{1/q} \leq \rho$.*

Assumption 2.4 posits the existence of a random variable $\boldsymbol{U}'$ whose average distance (as measured by the L^q norm) from the true vector of latent variables $\boldsymbol{U}$ is bounded above by some value $\rho \in \mathbb{R}_+$. The distribution of $\boldsymbol{U}'$ can be treated as a known baseline distribution for $\boldsymbol{U}$, in which case Assumption 2.4 restricts the true vector of latent variables to be "close" to that of $\boldsymbol{U}'$. Note Assumption 2.4 assumes there exists a random vector with a known distribution $P_{\boldsymbol{U}'}$, but does not assume the random vector $\boldsymbol{U}'$ itself is known or observed. In this way, Assumption 2.4 imposes no restrictions on the joint dependence between the random vector $\boldsymbol{U}'$ and $(\boldsymbol{Y}, \boldsymbol{Z})$, other than the dependence implied through the random vector $\boldsymbol{U}$.

The precise link between the distributions of $\boldsymbol{U}$ and $\boldsymbol{U}'$ implied by Assumption 2.4 is provided by the following result. Recall the Wasserstein distance between two probability measures Q_1 and Q_2 on $\mathbb{R}^{d_u}$ with finite q^{th} moments is given by:[8]

$$W_q(Q_1, Q_2) := \left(\inf_{\pi \in \Pi(Q_1, Q_2)} \int ||\boldsymbol{u}_1 - \boldsymbol{u}_2||_q^q \, d\pi \right)^{1/q} = \inf_{\tilde{\boldsymbol{U}} \sim Q_1, \tilde{\boldsymbol{U}}' \sim Q_2} (\mathbb{E}[||\tilde{\boldsymbol{U}} - \tilde{\boldsymbol{U}}'||_q^q])^{1/q}, \quad (2.15)$$

where $\Pi(Q_1, Q_2)$ denotes the set of all joint distributions on $\mathcal{U} \times \mathcal{U}$ with marginal distributions Q_1 and Q_2, and the infimum on the right side is over all couplings $(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{U}}')$ of the random vectors $(\boldsymbol{U}, \boldsymbol{U}')$. Furthermore, let $P_{\boldsymbol{U}}$ denote the distribution of $\boldsymbol{U}$ on $\mathcal{U}$ and let $P_{\boldsymbol{U}'}$ denote the distribution of $\boldsymbol{U}'$ on $\mathcal{U}$.

Proposition 2.3. *Suppose Assumption 2.1 holds.*

(i) If Assumption 2.4 holds, then $W_q(P_{\boldsymbol{U}}, P_{\boldsymbol{U}'}) \leq \rho$.

(ii) If $W_q(P_{\boldsymbol{U}}, P_{\boldsymbol{U}'}) \leq \rho$ then there exists a probability space $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ and random vectors $\tilde{\boldsymbol{U}} \sim \boldsymbol{U}$ and $\tilde{\boldsymbol{U}}' \sim \boldsymbol{U}'$ defined on $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ such that $(\tilde{\mathbb{E}}[||\tilde{\boldsymbol{U}} - \tilde{\boldsymbol{U}}'||_q^q])^{1/q} \leq \rho$; that is, Assumption 2.4 holds for $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{U}}'$.

[8] See Definition 6.2 in Villani (2009).

(iii) *If* $W_q(P_{\boldsymbol{U}}, P_{\boldsymbol{U}'}) \leq \rho$, *if* $\mathcal{U}$ *is a closed convex subset of* $\mathbb{R}^{d_u}$, *and if* $P_{\boldsymbol{U}}$ *admits a density with respect to the Lebesgue measure on* $\mathbb{R}^{d_u}$, *then there exists a random vector* $\boldsymbol{U}' \sim P_{\boldsymbol{U}'}$ *on* $(\Omega, \mathfrak{F}, \mathbb{P})$ *such that* $(\mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}'||_q^q])^{1/q} \leq \rho$; *that is, Assumption 2.4 holds for* $\boldsymbol{U}$ *and* $\boldsymbol{U}'$.

Parts (i) and (ii) of Proposition 2.3 follow directly from the definition of the Wasserstein distance in (2.15). Part (i) implies that Assumption 2.4 directly imposes a bound on the Wasserstein distance between $P_{\boldsymbol{U}}$ and $P_{\boldsymbol{U}'}$. Part (ii) implies that, if the Wasserstein distance between $P_{\boldsymbol{U}}$ and $P_{\boldsymbol{U}'}$ is bounded by ρ, then there exists a pair of random vectors with distributions $P_{\boldsymbol{U}}$ and $P_{\boldsymbol{U}'}$ satisfying Assumption 2.4. As a technical aside, the random vectors from part (ii) may not be defined on the initial probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, although part (iii) shows this can be guaranteed under some additional conditions. In fact, the conditions in part (iii) guarantees the existence of a measurable function $T : \mathcal{U} \to \mathcal{U}$ such that the pair $(\boldsymbol{U}, T(\boldsymbol{U}))$ solves the problem in (2.15) (see Champion and De Pascale (2011) Theorem 1.1).

Given the close connection between the constraint in Assumption 2.4 and the Wasserstein distance, we refer the constraint as the *q-Wasserstein penalty* or the *q-Wasserstein constraint*. In addition to acting as a metric between probability distributions, the Wasserstein distance has a straightforward interpretation in our setting: Assumption 2.4 allows the average distance between $\boldsymbol{U}$ and the baseline vector $\boldsymbol{U}'$ to be at most ρ.

Remark 2.7. *Assumption 2.4 can also be modified in a straightforward manner to impose the constraint on various subvectors of* $\boldsymbol{U}$. *This can be used to impose a baseline distribution for either only a few elements of* $\boldsymbol{U}$, *or for the entire vector* $\boldsymbol{U}$. *Furthermore, certain modifications of the assumption can be used to impose restrictions on the dependence between subvectors of* $\boldsymbol{U}$. *For instance, if* $\boldsymbol{U}' \sim N(\boldsymbol{0}, \boldsymbol{I})$ *and* $\rho = 0$, *then Assumption 2.4 and Proposition 2.3 imply* $\boldsymbol{U}$ *is equal to* $\boldsymbol{U}'$ *in distribution, and thus all subvectors of* $\boldsymbol{U}$ *are independent. If instead we assume* $U_k' \sim N(0,1)$ *for* $k = 1, \ldots, d_u$, *and modify Assumption 2.4 to impose* $(\mathbb{E}[||U_k - U_k'||_q^q])^{1/q} \leq \rho$ *for* $k = 1, \ldots, d_u$ *and* $\rho = 0$, *then Proposition 2.3 implies* U_k *is equal to* U_k' *in distribution, but the there are no restrictions on the dependence between* $U_1, \ldots, U_{d_u}$.

A revised version of Assumption 2.3 to accommodate the $q-$Wasserstein penalty is given by Assumption S.2.1 in Section S.2 of the Supplemental Material. Furthermore, Section S.2 discusses how Definition 2.1 can be revised to accommodate Assumptions 2.4 and S.2.1. After these changes, we have the following result.

Theorem 2.3. *Suppose Assumptions 2.1, 2.2, 2.4 and S.2.1 hold, and define:*

$$h_\rho(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}', \boldsymbol{u}, \boldsymbol{y}^\star, \mu, \boldsymbol{\lambda}, \theta) := \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) + \mu(||\boldsymbol{u} - \boldsymbol{u}'||_q^q - \rho^q) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta) \rangle.$$

Then $\overline{conv}(\Phi^*) = [\varphi_{\ell b}, \varphi_{ub}]$, *where:*

$$\varphi_{\ell b} = \inf_{\theta\in\Theta} \sup_{(\mu,\boldsymbol{\lambda})\in\mathbb{R}_+\times\mathbb{R}^{d_m}} \int \inf_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \inf_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} h_\rho(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U}',\boldsymbol{u},\boldsymbol{y}^\star,\mu,\boldsymbol{\lambda},\theta)\, d\mathbb{P},$$

$$\varphi_{ub} = \sup_{\theta\in\Theta} \inf_{(\mu,\boldsymbol{\lambda})\in\mathbb{R}_-\times\mathbb{R}^{d_m}} \int \sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} h_\rho(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U}',\boldsymbol{u},\boldsymbol{y}^\star,\mu,\boldsymbol{\lambda},\theta)\, d\mathbb{P}.$$

Theorem 2.3 demonstrates that Theorem 2.1 can be modified to accommodate the additional robustness constraint introduced in Assumption 2.4. The addition of the Wasserstein penalty makes our framework similar to approaches taken in the machine learning and operations research literature to deal with distributional robustness, especially Mohajerin Esfahani and Kuhn (2018), Sinha et al. (2018), and Blanchet and Murthy (2019).[9]

Both Assumption 2.4 and Theorem 2.3 hide an important practical issue: Assumption 2.4 imposes only that there is *some* random vector $\boldsymbol{U}'$ with the known distribution $P_{\boldsymbol{U}'}$ satisfying the $q-$Wasserstein constraint, although exactly which random vector satisfies the constraint may be unknown.[10] Practical use of Theorem 2.3 requires access to the random vector $\boldsymbol{U}'$ that satisfies the $q-$Wasserstein constraint from Assumption 2.4. In Section 3 we discuss practical implementation issues, and show how to implement the $q-$Wasserstein penalty by constructing $\boldsymbol{U}'$ using a sequence of discrete optimal transport problems.

2.4 Comparison to Christensen and Connault (2023)

In an important recent paper, Christensen and Connault (2023) also consider the problem of bounding scalar counterfactual quantities in structural models with a particular focus on exploring robustness to parametric distributional assumptions. Although inspired by their work, our framework has a number of differences with the framework of Christensen and Connault (2023). In this section we focus on describing two main differences: (i) the different expressions for the counterfactual bounds, and (ii) the different notions of robustness.

To ease comparison, we begin by summarizing the main features of the approach of Christensen and Connault (2023). The model in Christensen and Connault (2023) is defined using the following set of finite unconditional moments:

$$\mathbb{E}_{P_{\boldsymbol{U}}}[\boldsymbol{m}_1(\boldsymbol{U},\theta,\gamma_0)] \leq \boldsymbol{q}_{10}, \qquad \mathbb{E}_{P_{\boldsymbol{U}}}[\boldsymbol{m}_2(\boldsymbol{U},\theta,\gamma_0)] = \boldsymbol{q}_{20}, \tag{2.16}$$

where $\gamma_0 \in \Gamma$ is an auxiliary parameter belonging to a metric space Γ, and $\boldsymbol{q}_0 = (\boldsymbol{q}_{01}, \boldsymbol{q}_{02})$ is a vector

[9] Note that, in our notation, these papers are mostly concerned with robustness with respect to deviations from the empirical distribution of $(\boldsymbol{Y},\boldsymbol{Z})$ in prediction problems, and so are different from the problems the consider in this paper.

[10] For instance, if $\boldsymbol{U}_1', \boldsymbol{U}_2' \sim P_{\boldsymbol{U}'}$ this does not imply that $\boldsymbol{U}_1' = \boldsymbol{U}_2'$ a.s., and does not imply $\mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}_1'||] = \mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}_2'||]$. Because of this, it is possible $\mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}_1'||] \leq \rho < \mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}_2'||]$. This is because Wasserstein distances are not *law-invariant* in the terminology of Shapiro (2017).

of targeted moments that can depend on $(\boldsymbol{Y}, \boldsymbol{Z})$.[11] It is assumed the researcher has consistent estimates $(\hat{\boldsymbol{q}}, \hat{\gamma})$ of $(\boldsymbol{q}, \gamma_0)$, and the researcher is interested in a scalar counterfactual of the form $\mathbb{E}_{P_{\boldsymbol{U}}}[k(\boldsymbol{U}, \theta, \gamma_0)]$. The framework of Christensen and Connault (2023) allows researchers to assess sensitivity with respect to some baseline distribution $P_{\boldsymbol{U}'}$ for the latent variables by restricting $P_{\boldsymbol{U}}$ to belong to a neighborhood $\mathcal{N}_\rho$ of $P_{\boldsymbol{U}'}$. Following Hansen and Sargent (2001), Maccheroni et al. (2006), and Duchi and Namkoong (2021), the neighborhoods considered by Christensen and Connault (2023) are defined using a $\phi-$divergence. In particular, they consider the neighborhood $\mathcal{N}_\rho = \{P_{\boldsymbol{U}} : D_\phi(P_{\boldsymbol{U}} \mid P_{\boldsymbol{U}'}) \leq \rho\}$, where:

$$D_\phi(P_{\boldsymbol{U}} \mid P_{\boldsymbol{U}'}) = \begin{cases} \int \phi\left(\frac{dP_{\boldsymbol{U}}}{dP_{\boldsymbol{U}'}}\right) dP_{\boldsymbol{U}'}, & \text{if } P_{\boldsymbol{U}} \ll P_{\boldsymbol{U}'}, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\phi : [0, \infty) \to \mathbb{R}_+$ is a convex function differentiable on $(0, +\infty)$, satisfying some other regularity conditions.[12] Specific examples include the Kullback-Liebler, χ^2, and L^p-divergence. For a fixed tuple $(\theta, \gamma, \boldsymbol{q})$, the smallest value of the counterfactual parameter is given by:

$$\inf_{P_{\boldsymbol{U}} \in \mathcal{N}_\rho} \mathbb{E}_{P_{\boldsymbol{U}}}[k(\boldsymbol{U}, \theta, \gamma)], \quad \text{s.t. (2.16) holding at } (\theta, \gamma, \boldsymbol{q}).$$

Under some conditions, Christensen and Connault (2023) also use convex duality to show the problem can be rewritten as:[13]

$$\sup_{\eta > 0, \zeta \in \mathbb{R}, \boldsymbol{\lambda} \in \Lambda} -\eta \mathbb{E}_{P_{\boldsymbol{U}'}}\left[\phi^\star\left(\frac{k(\boldsymbol{U}', \theta, \gamma) + \zeta + \langle \boldsymbol{\lambda}, \boldsymbol{g}(\boldsymbol{U}', \theta, \gamma)\rangle}{-\eta}\right)\right] - \eta\rho - \zeta - \boldsymbol{\lambda}^\top \boldsymbol{q}, \tag{2.17}$$

where $\boldsymbol{g}(\,\cdot\,, \theta, \gamma)$ is a vector-valued function formed from stacking the moment functions from (2.16), and $\phi^\star$ is the convex conjugate of the function ϕ that defines the neighborhood $\mathcal{N}_\rho$. Under some additional assumptions, Christensen and Connault (2023) also show that when $\rho \to \infty$ the program (2.17) converges to the program:

$$\sup_{\boldsymbol{\lambda} \in \Lambda:\ \text{ess inf}(k(\,\cdot\,, \theta, \gamma) + \langle \boldsymbol{\lambda}, \boldsymbol{g}(\,\cdot\,, \theta, \gamma)\rangle) > -\infty} \text{ess inf}(k(\,\cdot\,, \theta, \gamma) + \langle \boldsymbol{\lambda}, \boldsymbol{g}(\,\cdot\,, \theta, \gamma)\rangle - \boldsymbol{\lambda}_{12}^\top \boldsymbol{q}), \tag{2.18}$$

where the essential infimum is with respect to the baseline measure $P_{\boldsymbol{U}'}$ (see Lemma B.1 and B.2 in Christensen and Connault (2023)).[14] Unlike our main results, the dual problem in (2.17) is derived using convex duality in infinite-dimensional spaces. In particular, the result relies on a dual pairing between Orlicz spaces, which in turn imposes restrictions on how the moment conditions and the

[11] When $\boldsymbol{Z}$ is discrete their framework also allows for conditional (on $\boldsymbol{Z} = \boldsymbol{z}$) moments, and moments that are non-separable between $\boldsymbol{U}$ and $\boldsymbol{Z}$: see Appendix A.2 and A.3 of Christensen and Connault (2023).

[12] See p. 272 of Christensen and Connault (2023).

[13] See Christensen and Connault (2023) Proposition 2.1.

[14] Recall the $P_{\boldsymbol{U}'}-$essential infimum of a measurable function $f : \mathcal{U} \to \mathbb{R}$ is defined as $\text{ess inf}(f) = \sup\{M : P_{\boldsymbol{U}'}(f(\boldsymbol{U}') < M) = 0\}$.

counterfactual functional can depend on the outcomes $\boldsymbol{Y}$ and the covariates $\boldsymbol{Z}$. These restrictions, among others, mean our results cannot be obtained as a special case of the results in Christensen and Connault (2023). Conversely, our approach is not amenable to neighborhoods that are constructed using $\phi-$divergences, which impose constraints on Radon-Nikodym derivatives. Thus, to the best of our knowledge, neither our approach nor the approach of Christensen and Connault (2023) is a special case of the other.

While the two approaches have a number of theoretical differences, there are important practical differences in how the two methods handle support-like restrictions on the latent and counterfactual variables. For this reason, it is useful to define the concept precisely.

Definition 2.2 (Support-Like Restrictions)**.** *We say the moment function* $f : \mathcal{Y}\times\mathcal{Z}\times\mathcal{U}\times\mathcal{Y}\times\Theta \to \mathbb{R}$ *(or, equivalently, the moment condition* $\mathbb{E}[f(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{Y}^\star,\theta)] = 0$*) imposes a support-like restriction if, for any* $\theta \in \Theta$*, we have* $\mathbb{E}[f(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{Y}^\star,\theta)] = 0$ *if and only if* $f(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{Y}^\star,\theta) = 0$ *a.s.*

For instance, the conditions $\mathbb{E}[\delta_1(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\theta)] = 0$ and $\mathbb{E}[\delta_2(\boldsymbol{Y}^\star,\boldsymbol{Z},\boldsymbol{U},\theta)] = 0$ both impose support-like restrictions in the primal problem, restricting the values of $\boldsymbol{U}$ and $\boldsymbol{Y}^\star$ using random set theory. Similar to the previous insights of Ekeland et al. (2010), Beresteanu et al. (2011), and Li (2021), our approach treats these conditions different from the other moment conditions. For Christensen and Connault (2023) to maintain sharpness the identifying information contained in the support-like restrictions must be captured by an equivalent set of moment conditions of the form (2.16). In contrast, we are able to avoid imposing a large number of moment conditions by absorbing all support-like restrictions on the latent and counterfactual variables inside the inner optimization problems in (2.13) and (2.14). This introduces an additional optimization problem in our formulation, but otherwise reduces the dimension of the middle problem over $\boldsymbol{\lambda} \in \mathbb{R}^{d_m}$.[15]

For models that are challenging to describe using a small number of moment conditions, we believe our method has an advantage. This is the case when the covariates have infinite but countable support, the model is incomplete, or the counterfactual model is incomplete. For instance, in models of strategic interaction (e.g. Example 2 and Example 3) the number of moment conditions grows exponentially in the number of interacting agents, and constructing the moment conditions requires an expensive simulation procedure that enumerates all equilibria for each draw of the latent variables.[16] The difficulty of describing incomplete models using moment conditions also often forces researchers to use non-sharp characterizations (see the discussion in Kédagni et al. (2020)). The advantages of using random set theory to provide sharp characterizations in these environments has been discussed at length in Section 3 of Beresteanu et al. (2011). When the counterfactual model is also incomplete (e.g. Example 2 and Example 3), bounding marginal

[15] Since the conditions $\mathbb{E}[\delta_1(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{U},\theta)] = 0$ and $\mathbb{E}[\delta_2(\boldsymbol{Y}^\star,\boldsymbol{Z},\boldsymbol{U},\theta)] = 0$ are not separable in $(\boldsymbol{Y},\boldsymbol{Z})$, they cannot be written as special cases of the moment conditions in (2.16).

[16] See Section 3 in Sheng (2020) for a discussion of the issue in the case of a network formation game.

effects or other counterfactual parameters in models described by moment conditions is also known to be difficult (see Ciliberto and Tamer (2009) pp. 1820-1822). To illustrate some of these points concretely, we revisit Example 1 (discrete choice) and Example 2 (binary game) in Section S.3.1 of the Supplemental Material, and provided a side-by-side comparison between our approach and the approach taken in Christensen and Connault (2023).

There are also some benefits to our notion of robustness versus the notion of robustness considered in Christensen and Connault (2023).[17] Some of the benefits are related to interpretation. As explained in the previous section, the Wasserstein distance has a relatively simple interpretation, with the value of ρ acting as an upper bound on the average distance between $\boldsymbol{U}$ and $\boldsymbol{U}'$ for some random vector $\boldsymbol{U}'$ with a known distribution. The Wasserstein distance is also a metric between probability distributions, metrizing weak convergence (see Villani (2009) Theorem 6.9). In contrast, divergences have challenging interpretations and are not metrics, since they need not be symmetric or satisfy the triangle inequality. Our use of Wasserstein distances also allows for distributions of $\boldsymbol{U}$ that are discrete, which are ruled out by divergences when the reference measure $P_{\boldsymbol{U}'}$ is continuous. Finally, Wasserstein distances capture the metric properties of the underlying probability space: distributions on $\mathcal{U}$ that are close in terms of the norm on $\mathcal{U}$ also tend to have a small Wasserstein distance. In contrast, divergences rely on pointwise comparisons of the supports of two distributions, without consideration of the geometry of the space on which they are supported. This property can produce counter-intuitive examples: two distributions that appear to be close can have arbitrarily large divergence, and two distributions that appear to be far can have the same divergence as two distributions that appear to be close. The fact the Wasserstein distance reflects the underlying geometry of the latent variable space is especially useful when the magnitude or numerical values of the latent variables is meaningful, as is the case, for example, when the latent variables represent measurement error or unobserved components of profit.[18] A similar debate between divergences and the Wasserstein distance also exists in the distributional robustness literature (see Wozabal (2012) and Blanchet and Murthy (2019)).

Despite the differences between the two approaches, some connections can be made between our approach and the nonparametric bounds (i.e. when $\rho \to +\infty$) of Christensen and Connault (2023). In Section S.3.2 of the Supplemental Material we show the nonparametric bounds of Christensen and Connault (2023) presented in (2.18) can be obtained as a special case of the bounds presented in Theorem 2.1 and Theorem 2.2 under some additional assumptions.

[17] See Chapter 8 in Peyré et al. (2019) for an overview.
[18] See Schennach and Starck (2022) for a similar point.

$$\hat{\varphi}_{\ell b} := \underbrace{\inf_{\theta\in\Theta} \underbrace{\sup_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \overbrace{\inf_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \inf_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \left(\varphi(\boldsymbol{y}^\star,\boldsymbol{z}_i,\boldsymbol{u},\theta) + \langle\boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{z}_i,\boldsymbol{u},\theta)\rangle\right)}^{\text{Local Problem: } \varphi^L_{\ell b}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta,\boldsymbol{\lambda})}}_{\text{Average Local Problem: } \hat{\varphi}^L_{\ell b}(\theta,\boldsymbol{\lambda})}}_{\text{Middle Problem: } \hat{\varphi}^M_{\ell b}(\theta)}}_{\text{Outer Problem: } \hat{\varphi}_{\ell b}}.$$

Figure 2: The nested optimization defining the sample analog lower bound.

3 Estimation

In this section we consider estimation of the lower and upper endpoints in Theorem 2.2 with their corresponding sample averages. With the exception of the local problems, the algorithms discussed in this section are general-purpose and do not rely on any special structure arising in particular problems. We focus our discussion on estimating $\varphi_{\ell b}$, as the corresponding estimator for φ_{ub} is similar. To proceed, we decompose the lower bound into a sequence of nested optimization problems, illustrated in Figure 2. For simplicity, we exclude the random vector $\boldsymbol{U}'$ from Section 2.3 in most of the notation except in the examples of Section 3.1 and later in Section 3.3, which explicitly deals with computational issues associated with the Wasserstein constraints.

3.1 The Local Problem

The exact form of the local problem is application-specific, but many examples can be formulated as an LP, or an MILP. Efficient algorithms exist for both LPs and MILPs, and so the local problem can typically be solved to global optimality within a small fraction of a second. We revisit the examples from Section 2.1 to illustrate how this can be done. In each of the examples, we impose a $1-$Wasserstein penalty assuming the realizations of the random vector $\boldsymbol{U}'$ from Assumption 2.4 are observed. We return to the problem of constructing realizations of $\boldsymbol{U}'$ in Section 3.3.

Example 1 (Multinomial Choice, Cont'd)**.** *Consider again the multinomial choice example where consumers choose among J alternatives. The utility received from alternative j is given by:*

$$\pi_j(\boldsymbol{z}_i,\boldsymbol{u}_i) = \boldsymbol{z}_{ij}^\top\theta_j + u_{ij}.$$

Consider a counterfactual where $\boldsymbol{z}_i$ is replaced by $\check{\boldsymbol{z}}_i = \check{\boldsymbol{z}}_i(\boldsymbol{z}_i)$, and suppose the objective is to compute a lower bound on the change in social surplus:

$$\mathbb{E}\left[\max_{j\in\mathcal{J}}\{\check{\boldsymbol{z}}_{ij}^\top\theta_j + u_{ij}\} - \max_{j\in\mathcal{J}}\{\boldsymbol{z}_{ij}^\top\theta_j + u_{ij}\}\right],$$

while imposing a $1-$*Wasserstein penalty between* $\boldsymbol{u}_i$ *and a vector* $\boldsymbol{u}'_i$ *with a known distribution. Given the observed choice* y_i*, the vector of latent variables* $\boldsymbol{u}_i$ *satisfies:*

$$\boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{iy_i} \geq \boldsymbol{z}_{ij}^\top \theta_j + u_{ij}, \tag{3.1}$$

for $j = 1, \ldots, J$*. Rewriting the objective function, we have:*

$$\varphi(\boldsymbol{y}_i^\star, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \max_{j \in \mathcal{J}} \left\{ \check{\boldsymbol{z}}_{ij}^\top \theta_j + u_{ij} \right\} - \max_{j \in \mathcal{J}} \left\{ \boldsymbol{z}_{ij}^\top \theta_j + u_{ij} \right\} = \max_{j \in \mathcal{J}} \left\{ \check{\boldsymbol{z}}_{ij}^\top \theta_j - \boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{ij} - u_{iy_i} \right\}.$$

Now introducing the $1-$*Wasserstein penalty, the objective function for the local problem becomes:*

$$\max_{j \in \mathcal{J}} \left\{ \check{\boldsymbol{z}}_{ij}^\top \theta_j - \boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{ij} - u_{iy_i} \right\} + \mu \left(\sum_{j=1}^{J} |u_{ij} - u'_{ij}| - \rho \right). \tag{3.2}$$

To simplify, let $\check{u}_{ij}$ *be a variable satisfying:*

$$\check{u}_{ij} \geq u_{ij} - u'_{ij}, \qquad\qquad \check{u}_{ij} \geq u'_{ij} - u_{ij}, \tag{3.3}$$

for $j = 1, \ldots, J$*, and let* t *be a variable satisfying:*

$$t \geq \check{\boldsymbol{z}}_{ij}^\top \theta_j - \boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{ij} - u_{iy_i}, \tag{3.4}$$

for $j = 1, \ldots, J$*. Then* (3.2) *can be rewritten as:*

$$\max_{j \in \mathcal{J}} \left\{ \check{\boldsymbol{z}}_{ij}^\top \theta_j - \boldsymbol{z}_{iy_i}^\top \theta_{y_i} + u_{ij} - u_{iy_i} \right\} + \mu \left(\sum_{j=1}^{J} |u_{ij} - u'_{ij}| - \rho \right) = \min_{t \in \mathbb{R}} t + \mu \left(\sum_{j=1}^{J} \check{u}_{ij} - \rho \right),$$

subject to the constraints (3.3) *and* (3.4)*. Combining everything the local problem can be written as:*

$$\min_{\boldsymbol{u}, \check{\boldsymbol{u}} \in \mathbb{R}^J} \min_{t \in \mathbb{R}} t + \mu \left(\sum_{j=1}^{J} \check{u}_{ij} - \rho \right), \tag{3.5}$$

subject to the constraints (3.1)*,* (3.3)*, and* (3.4)*. Note the objective function is linear in* $(\boldsymbol{u}, \check{\boldsymbol{u}}, t)$*, and the constraints* (3.1)*,* (3.3)*, and* (3.4) *are all linear, so* (3.5) *is an LP with* $2J+1$ *variables and* $4J$ *constraints.*

Example 2 (Binary Game, Cont'd)**.** *Consider again the game with* $K = 2$ *players where payoffs in the* i^{th} *game are given by:*

$$\pi_{i1}(\boldsymbol{y}, \boldsymbol{z}; \theta) := y_{i1}(\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 - u_{i1}), \qquad \pi_{i2}(\boldsymbol{y}, \boldsymbol{z}; \theta) := y_{i2}(\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 - u_{i2}).$$

Under the assumption of pure strategy Nash equilibria the observed entry decisions satisfy:

$$y_{i1} = \mathbb{1}\{\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 \geq u_{i1}\}, \qquad y_{i2} = \mathbb{1}\{\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 \geq u_{i2}\}. \tag{3.6}$$

Now consider a counterfactual that replaces $(\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ *with* $(\check{\boldsymbol{z}}_{i1}, \check{\boldsymbol{z}}_{i2})$ $\forall i$*, and suppose the researcher wishes to study the effect of the intervention on player 1's entry probability while imposing a* 1*-Wasserstein penalty. Then the realized counterfactual outcomes* $(y_{i1}^\star, y_{i2}^\star)$ *satisfy:*

$$y_{i1}^\star = \mathbb{1}\{\check{\boldsymbol{z}}_{i1}^\top \beta_1 + y_{i2}^\star\delta_1 \geq u_{i1}\}, \qquad y_{i2}^\star = \mathbb{1}\{\check{\boldsymbol{z}}_{i2}^\top \beta_2 + y_{i1}^\star\delta_2 \geq u_{i2}\}. \tag{3.7}$$

Furthermore, imposing the 1-Wasserstein penalty, the researcher's local objective function can be written as:

$$y_{i1}^\star + \mu\left(\sum_{k=1}^{2} \check{u}_{ik} - \rho\right), \tag{3.8}$$

where $(\check{u}_{i1}, \check{u}_{i2})$ *are auxiliary variables satisfying the constraints* $\check{u}_{ik} \geq u_{ik} - u'_{ik}$ *and* $\check{u}_{ik} \geq u'_{ik} - u_{ik}$*. For a large value* M *and a small value* $\eta > 0$*, the constraints* (3.6) *can written as:*

$$\begin{aligned} -M(1 - y_{i1}) \leq \boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 - u_{i1} \leq My_{i1} - \eta, \\ -M(1 - y_{i2}) \leq \boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 - u_{i2} \leq My_{i2} - \eta. \end{aligned}$$

In addition, the constraints (3.7) *can be written as:*

$$\begin{aligned} -M(1 - y_{i1}^\star) \leq \boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}^\star\delta_1 - u_{i1} \leq My_{i1}^\star - \eta, \\ -M(1 - y_{i2}^\star) \leq \boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}^\star\delta_2 - u_{i2} \leq My_{i2}^\star - \eta, \end{aligned}$$

plus the integer constraints $y_{i1}^\star, y_{i2}^\star \in \{0, 1\}$*. Defining the optimizing variable* $\boldsymbol{w}_i = (\boldsymbol{u}_i, \boldsymbol{y}_i^\star, \check{\boldsymbol{u}}_i)$*, where* $\boldsymbol{u}_i = (u_{i1}, u_{i2})$*,* $\boldsymbol{y}_i = (y_{i1}^\star, y_{i2}^\star)$*, and* $\check{\boldsymbol{u}}_i = (\check{u}_{i1}, \check{u}_{i2})$*, all constraints above can be expressed as* $\boldsymbol{y}^\star \in \{0, 1\}^2$ *plus* $\boldsymbol{A}\boldsymbol{w}_i \leq \boldsymbol{b}$*. Maximizing* (3.8) *subject to these constraints is an MILP. In the general case with* K *players, the MILP has* $3K$ *variables (of which* K *are integer variables) and* $4K$ *inequality constraints.*

Example 3 (Network Formation, Cont'd)**.** *Consider a network formation game with* $K = 3$ *agents, where the marginal utilities of link formation between agent* k *and* k' *are given by:*

$$\Delta\pi_{kk'}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \theta_1 + \boldsymbol{z}_{ik}^\top \theta_2 + ||\boldsymbol{z}_{ik} - \boldsymbol{z}_{ik'}||\theta_3 + \theta_4 y_{ik'k''} + u_{ikk'},$$

where $k'' \neq k, k'$*. Under pairwise stability and nontransferable utility the observed links satisfy:*

$$y_{ikk'} = \mathbb{1}\{\Delta\pi_{kk'}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\}\mathbb{1}\{\Delta\pi_{k'k}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \geq 0\}, \tag{3.9}$$

for all $k \neq k'$. Now consider a counterfactual that replaces $\boldsymbol{z}_i$ with $\check{\boldsymbol{z}}_i = \check{\boldsymbol{z}}_i(\boldsymbol{z}_i)$, and suppose the researcher wishes to study the effect of the intervention on the average number of links between agents of type $\mathcal{T}_1(\boldsymbol{z}_i) \subset \{1,2,3\}$ with agents of type $\mathcal{T}_2(\boldsymbol{z}_i) \subset \{1,2,3\}$, where $\mathcal{T}_1(\boldsymbol{z}_i) \cap \mathcal{T}_2(\boldsymbol{z}_i) = \varnothing$. The counterfactual outcomes $\boldsymbol{y}_i^\star = (y_{i12}^\star, y_{i13}^\star, y_{i23}^\star)$ satisfy:

$$y_{ikk'}^\star = \mathbb{1}\{\Delta\pi_{kk'}(\boldsymbol{y}_i^\star, \check{\boldsymbol{z}}_i, \boldsymbol{u}_i, \theta) \geq 0\}\mathbb{1}\{\Delta\pi_{k'k}(\boldsymbol{y}_i^\star, \check{\boldsymbol{z}}_i, \boldsymbol{u}_i, \theta) \geq 0\}, \tag{3.10}$$

for all $k \neq k'$. Furthermore, imposing the 1-Wasserstein penalty, the researcher's local objective function can be written as:

$$\sum_{k \in \mathcal{T}_1(\boldsymbol{z}_i)} \sum_{k' \in \mathcal{T}_2(\boldsymbol{z}_i)} y_{ikk'}^\star + \mu \left(\sum_{k \neq k'} \check{u}_{ikk'} - \rho \right), \tag{3.11}$$

where $\check{u}_{ikk'}$ are auxiliary variables satisfying the constraints $\check{u}_{ikk'} \geq u_{ikk'} - u'_{ikk'}$ and $\check{u}_{ikk'} \geq u'_{ikk'} - u_{ikk'}$. For a large value M and a small value $\eta > 0$, the constraints (3.9) can be represented by the linear constraints:

$$y_{ikk'} \leq s_{ikk'}, \qquad y_{ikk'} \leq s_{ik'k}, \qquad y_{kk'} \geq s_{ikk'} + s_{ik'k} - 1,$$

where $s_{ikk'}$ and $s_{ik'k}$ satisfy:

$$\begin{aligned} -M(1 - s_{ikk'}) &\leq \Delta\pi_{kk'}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \leq M s_{ikk'} - \eta, \\ -M(1 - s_{ik'k}) &\leq \Delta\pi_{k'k}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta) \leq M s_{ik'k} - \eta. \end{aligned}$$

These constraints impose all restrictions on the latent variable vector $\boldsymbol{u}_i$ implied by the observed vector $(\boldsymbol{y}_i, \boldsymbol{z}_i)$. In addition, the constraints (3.10) can be represented by the constraints:

$$y_{ikk'}^\star \in \{0,1\}, \qquad y_{ikk'}^\star \leq t_{ikk'}, \qquad y_{ikk'}^\star \leq t_{ik'k}, \qquad y_{ikk'}^\star \geq t_{ikk'} + t_{ik'k} - 1,$$

where $t_{ikk'}$ and $t_{ik'k}$ satisfy:

$$\begin{aligned} -M(1 - t_{ikk'}) &\leq \Delta\pi_{kk'}(\boldsymbol{y}_i^\star, \check{\boldsymbol{z}}_i, \boldsymbol{u}_i, \theta) \leq M t_{ikk'} - \eta, \\ -M(1 - t_{ik'k}) &\leq \Delta\pi_{k'k}(\boldsymbol{y}_i^\star, \check{\boldsymbol{z}}_i, \boldsymbol{u}_i, \theta) \leq M t_{ik'k} - \eta. \end{aligned}$$

Now define the optimizing variable $\boldsymbol{w}_i = (\boldsymbol{u}_i, \boldsymbol{y}_i^\star, \check{\boldsymbol{u}}_i, \boldsymbol{s}_i, \boldsymbol{t}_i)$, where $\boldsymbol{u}_i = (u_{i12}, u_{i21}, \ldots, u_{i32})$, $\boldsymbol{y}_i^\star = (y_{i12}^\star, y_{i13}^\star, y_{23}^\star)$, $\check{\boldsymbol{u}}_i = (\check{u}_{i12}, \check{u}_{i21}, \ldots, \check{u}_{i32})$, $\boldsymbol{s}_i = (s_{i12}, s_{i21}, \ldots, s_{i32})$ and $\boldsymbol{t}_i = (t_{i12}, t_{i21}, \ldots, t_{i32})$. All constraints can be expressed as $\boldsymbol{y}_i^\star \in \{0,1\}^3$ plus $\boldsymbol{A}\boldsymbol{w}_i \leq \boldsymbol{b}$. Maximizing (3.11) subject to these constraints is an MILP. In the general case with K agents, the MILP has $9K(K-1)/2$ variables (of which $K(K-1)/2$ are integer variables) and $6K(K-1)$ inequality constraints.

Although the examples use some special structure from each problem, a simple reformulation

of the local problem as an LP or an MILP appears to be typical for nonlinear models with discrete outcomes.

Remark 3.1. *Computation of the average local problem $\hat{\varphi}^L_{\ell b,n}(\theta, \boldsymbol{\lambda})$ can be simplified substantially when the supports $\mathcal{Y}$ and $\mathcal{Z}$ of the random vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are finite. In particular, when there are many repeated vectors $(\boldsymbol{y}_i, \boldsymbol{z}_i)$, computation time can be saved by solving each local problem only once at each unique pair $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ in the sample. The average local problem in Figure 2 can then be constructed by weighting by the relative frequency of each unique pair $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ in the sample. In this case the number of local problems to solve is at most $|\mathcal{Y}| \times |\mathcal{Z}|$, regardless of the sample size.*

3.2 The Middle Problem

The middle problem in Figure 2 accepts a fixed value of $\theta \in \Theta$, and maximizes the sample average local problem $\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda})$. If the local problems can be solved using LPs, in some cases linear programming duality results can be applied to solve both the middle and local problems simultaneously as a large linear or quadratic program. In the general case the objective function for the middle problem is concave in $\boldsymbol{\lambda} \in \Lambda$, although it is not always differentiable. However, we show the function $\hat{\varphi}^L_{\ell b}$ is sub-differentiable under some assumptions. In the following result, $ri(S)$ denotes the relative interior of the set S, and $\partial f(\overline{\boldsymbol{x}})$ denotes the subdifferential of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\bar{\boldsymbol{x}} \in \mathbb{R}^d$. Section S.1 of the Supplemental Material provides a formal definition.

Lemma 3.1. *Suppose Assumptions 2.1 and 2.2 hold and fix any $\theta \in \Theta$. Then the subdifferential of the function $\boldsymbol{\lambda} \mapsto -\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda})$ is nonempty, convex and closed at every $\boldsymbol{\lambda} \in ri(dom(-\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda})))$. Furthermore, suppose the set $\mathcal{S}_i := \{(\boldsymbol{u}, \boldsymbol{y}^\star) \in \mathcal{U} \times \mathcal{Y} : \boldsymbol{u} \in \mathcal{U}(\boldsymbol{y}_i, \boldsymbol{z}_i, \theta), \boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{u}, \boldsymbol{z}_i, \theta)\}$ is compact for $i = 1, \ldots, n$. Let $(\boldsymbol{u}^*_i, \boldsymbol{y}^{\star *}_i)$ denote a (possibly non-unique) optimal solution to the i^{th} local problem at some finite $\bar{\boldsymbol{\lambda}}$, and let:*

$$\boldsymbol{g}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}^*_i, \theta). \tag{3.12}$$

Then $-\boldsymbol{g}(\theta) \in \partial_{\boldsymbol{\lambda}}(-\hat{\varphi}^L_{\ell b}(\theta, \bar{\boldsymbol{\lambda}}))$; that is, $-\boldsymbol{g}(\theta)$ is a subgradient of the negative sample average local problem at $\bar{\boldsymbol{\lambda}}$. Furthermore, if the solution $(\boldsymbol{u}^*_i, \boldsymbol{y}^{\star *}_i)$ is unique for $i = 1, \ldots, n$, then $-\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda})$ is differentiable at $\bar{\boldsymbol{\lambda}}$, and $-\boldsymbol{g}(\theta) = -\nabla_{\boldsymbol{\lambda}} \hat{\varphi}^L_{\ell b}(\theta, \bar{\boldsymbol{\lambda}})$.

Lemma 3.1 provides a subgradient of the negative sample average local problem, and also provides conditions under which the subgradient is a true gradient. Provided at least one subgradient is known, there exists efficient algorithms for computing global minima of convex functions. Similar to gradient descent algorithms, these algorithms take small steps in the direction of the negative subgradient at each iteration. These algorithms use first-order information only, and each iteration is simple and fast.

Computing the subgradient in (3.12) requires solving all local problems. From the previous section, each local problem is simple to solve, but when n is large solving all the local problems to compute the subgradient in (3.12) at each step can be expensive.[19] To reduce computation time, we use a *stochastic subgradient descent (SSGD) algorithm* to solve the middle problem by evaluating the subgradient in (3.12) and taking a step using only a single observation at each iteration.[20] Generally speaking, these algorithms minimize a function by taking steps in the direction of the negative subgradient $-\boldsymbol{g}_{i_k} := -\boldsymbol{m}(\boldsymbol{y}_{i_k}, \boldsymbol{z}_{i_k}, \boldsymbol{u}^*_{i_k}, \theta)$ associated with the randomly sampled observation $i_k \in \{1, \ldots, n\}$ at each iteration. By construction, $-\boldsymbol{g}_{i_k}$ is an unbiased estimator of the true subgradient from (3.12), so SSGD algorithms take many noisy steps in directions that are unbiased for the true descent direction. Although the procedure requires a high number of iterations, the descent direction at each iteration is inexpensive to compute, and overall the procedure is much faster than a procedure that computes the full subgradient using (3.12) on each iteration (see Bottou et al. (2018)).

We benchmark four SSGD algorithms in our application in Section 5: the adaptive gradient descent (AdaGrad) algorithm of Duchi et al. (2011), the AdaDelta algorithm of Zeiler (2012), the RMSprop algorithm of Tieleman and Hinton (2012), and the adaptive moment estimation (Adam) algorithm of Kingma and Ba (2014). The general structure of the four algorithms is displayed in Algorithm 1 in Section S.4 of the Supplemental Material, with each algorithm differing in the step size used to update $\boldsymbol{\lambda}_k$ at each iteration given the randomly sampled subgradient $\boldsymbol{g}_{i_k}$.

In our final implementation we modify Algorithm 1 in a few ways, and include a coarse initial grid search to find a starting value $\boldsymbol{\lambda}_0$, and every few thousand iterations we use the full sample to compute $\hat{\varphi}^L_{\ell b}(\boldsymbol{\lambda}_k, \theta)$ in order to construct a stopping criterion. Full details are provided in Section S.4 in the Supplemental Material.

3.3 On Implementing the $q-$Wasserstein Constrained Bounds

Examples 1, 2, and 3 in Section 3.1 all impose a $1-$Wasserstein penalty, which requires access to realizations of a random vector $\boldsymbol{U}'$ satisfying Assumption 2.4. To construct the realizations in practice, we use an approach similar to the iterative match-and-update procedure of Arellano and Bonhomme (2021). Throughout, fix the value of $(\mu, \boldsymbol{\lambda})$. We first take n draws $\{\boldsymbol{u}'_i\}_{i=1}^n$ from the distribution $P_{\boldsymbol{U}'}$ using quasi-random Halton sequences. With the initial draws fixed, we solve all n local problems and collect the optimal solutions $\{\boldsymbol{u}^*_i\}_{i=1}^n$. We then solve an optimal transport problem to minimize the $q-$Wasserstein distance between the empirical distributions $n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{u}'_i}$ and $n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{u}^*_i}$, where

[19] Similar to Remark 3.1, if the data are discrete then the problem can be mitigated by computing the subgradient at each unique pair $(\boldsymbol{y}_i, \boldsymbol{z}_i)$ in the sample, and then averaging the resulting subgradients using the sample frequencies. Here we are concerned mostly with the general case with at least one variable with infinite support.

[20] See Bottou et al. (2018) for a review.

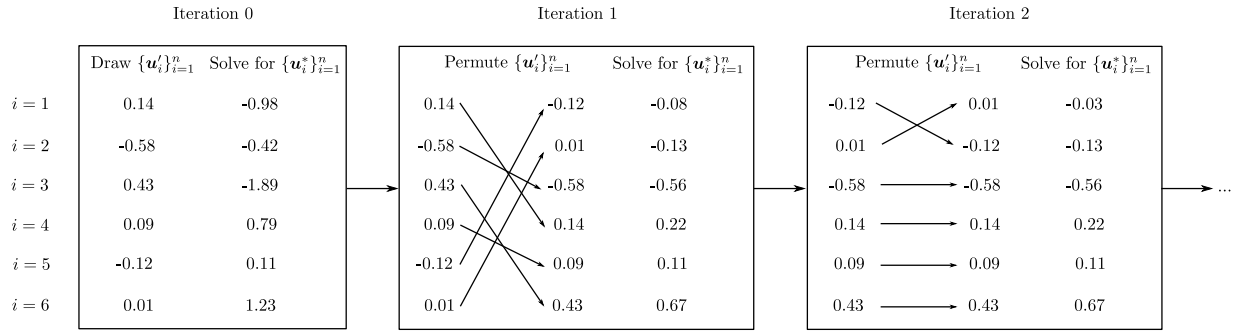


Figure 3: An illustration of the permutation procedure used to match the latent variable draws $\{\boldsymbol{u}'_i\}_{i=1}^n$ to the optimal values $\{\boldsymbol{u}^*_i\}_{i=1}^n$ from the local problem. The illustration shows a problem with $n = 6$ observations. Before the first iteration, the values $\{\boldsymbol{u}'_i\}_{i=1}^n$ are drawn from the distribution $P_{\boldsymbol{U}'}$. Using these values, initial values $\{\boldsymbol{u}^*_i\}_{i=1}^n$ are obtained by solving the local problems. At iteration $k \geq 1$ the initial draws $\{\boldsymbol{u}'_i\}_{i=1}^n$ are permuted to minimize the Wasserstein distance with the most recent values $\{\boldsymbol{u}^*_i\}_{i=1}^n$. After the permutation the values $\{\boldsymbol{u}^*_i\}_{i=1}^n$ are updated again, and procedure continues.

$\delta_{\boldsymbol{x}_i}$ denotes the Dirac-delta function.[21] The solution to the optimal transport problem determines a permutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ that minimizes $n^{-1} \sum_{i=1}^n |\boldsymbol{u}'_{\sigma(i)} - \boldsymbol{u}^*_i|^q$, which can then be used to update the points $\{\boldsymbol{u}'_i\}_{i=1}^n \leftarrow \{\boldsymbol{u}'_{\sigma(i)}\}_{i=1}^n$. With the updated points $\{\boldsymbol{u}'_i\}_{i=1}^n$, we then re-solve all n local problems to get updated solutions $\{\boldsymbol{u}^*_i\}_{i=1}^n$, and the procedure repeats. An illustration for the case when $d_u = 1$ and $n = 6$ is provided in Figure 3.

By construction, each iteration of the procedure weakly decreases the value of the average of the local problems, since each permutation step decreases the value of the current Wasserstein distance, and each optimization step decreases the value of each local problem. Termination occurs when the change in the empirical transport cost $n^{-1} \sum_{i=1}^n |\boldsymbol{u}'_{\sigma(i)} - \boldsymbol{u}^*_i|^q$ between successive iterations is below some threshold. In practice convergence is fast, with the empirical transport cost typically converging to within six decimals in 6 to 10 iterations in the application in Section 5. Since the procedure requires the evaluation of all n local problems, in practice it is interwoven with the middle problem procedure in the previous section, and is repeated every few thousand SSGD iterations when we check for convergence. Additional details on our implementation are provided in Section S.4 of the Supplemental Material.

3.4 The Outer Problem

Unlike the local and middle problems, the outer problem typically has limited structure that can be exploited during optimization. Furthermore, evaluating the middle problem at a fixed value of $\theta \in \Theta$ is relatively expensive. Because of these issues, we recommend a *response-surface method* to solve the outer problem. Response surface methods require a limited number of function evaluations, which is preferable when the objective function is expensive to evaluate (see Jones et al. (1998) p.

[21] When $d_u = 1$ this can be done by simply sorting the elements $\{\boldsymbol{u}_i\}_{i=1}^n$ and $\{\boldsymbol{u}'_i\}_{i=1}^n$, and when $d_u > 1$ this can be done by solving an LP.

457). These methods are able to optimize expensive black-box function by evaluating the function at a small number of points, fitting a flexible interpolating surface between the evaluation points, and then by choosing the next point of evaluation in a principled manner using the interpolating surface. Iterating on this procedure, response surface methods can achieve global convergence under weak assumptions.[22] We use an implementation of a response surface method described in Jones et al. (1998) called the *efficient global optimization* (EGO) algorithm, which uses a Gaussian process regression model for the response surface. The procedure is closely related to *Bayesian global optimization*, and was recently adapted by Kaido et al. (2019) for the problem of subvector inference for partially identified models.

Recall from Figure 2 that the outer problem minimizes the value function $\hat{\varphi}^M_{\ell b}(\theta)$ from the middle problem. Our outer problem algorithm begins by evaluating $\hat{\varphi}^M_{\ell b}(\theta)$ at an initial set of evaluation points $\theta_1, \ldots, \theta_L$ from Θ which are drawn using latin hypercube sampling. Following Kaido et al. (2019), we set $L = 10d_\theta + 1$, so the number of initial evaluation points grows linearly with the dimension of the parameter space. We then fit a Gaussian process regression model to the points $\{(\theta_\ell, \hat{\varphi}^M_{\ell b}(\theta))\}_{\ell=1}^L$. The Gaussian process regression model specifies a Gaussian process prior $GP(\mu(\cdot), K(\cdot,\cdot))$ for $\hat{\varphi}^M_{\ell b}(\cdot)$, where $\mu : \Theta \to \mathbb{R}$ is the mean function and $K : \Theta \times \Theta \to \mathbb{R}$ is a covariance kernel. In the application we take μ as a constant, and use a separable squared exponential kernel, which depends on a vector of hyperparameters $\boldsymbol{\beta}$. Using an empirical Bayes framework, estimates of the parameters μ, σ^2, and $\boldsymbol{\beta}$ are obtained via maximum likelihood. For given parameters μ, σ^2, $\boldsymbol{\beta}$, the posterior distribution of $\hat{\varphi}^M_{\ell b}(\cdot) \mid \{(\theta_\ell, \hat{\varphi}^M_{\ell b}(\theta_\ell))\}_{\ell=1}^L$ is a multivariate normal, and marginalizing the likelihood with respect to the posterior the predictive distribution at a new value $\theta \in \Theta$ is $N(\hat{\hat{\varphi}}^M_{\ell b}(\theta), \hat{\hat{\sigma}}^2_{\ell b}(\theta))$ where:

$$\hat{\hat{\varphi}}^M_{\ell b}(\theta) := \mu + \boldsymbol{k}(\theta)^\top \boldsymbol{K}^{-1}(\hat{\boldsymbol{\varphi}}^M - \mathbf{1}\mu), \qquad \hat{\hat{\sigma}}^2_{\ell b}(\theta) := K(\theta,\theta) - \boldsymbol{k}(\theta)^\top \boldsymbol{K}^{-1}\boldsymbol{k}(\theta), \tag{3.13}$$

where $\boldsymbol{K}$ is an $L \times L$ matrix with typical entry $K(\theta_\ell, \theta_{\ell'})$, $\hat{\boldsymbol{\varphi}}^M = (\hat{\varphi}^M_{\ell b}(\theta_1), \ldots, \hat{\varphi}^M_{\ell b}(\theta_L))^\top$, $\boldsymbol{k}(\theta) := (K(\theta,\theta_1), \ldots, K(\theta,\theta_L))^\top$, and where $\mathbf{1}$ is an $L \times 1$ vector of 1's.[23] The predictive mean $\hat{\hat{\varphi}}^M_{\ell b}(\theta)$ is used for the response surface model of the function $\hat{\varphi}^M_{\ell b}(\theta)$, and the predictive variance $\hat{\hat{\sigma}}^2_{\ell b}(\theta)$ is used as a model of uncertainty about the values of $\hat{\varphi}^M_{\ell b}(\theta)$.

The benefit of using a Gaussian process regression model for response surface optimization is twofold. First, the predictive mean in (3.13) interpolates the evaluation points $\{(\theta_\ell, \hat{\varphi}^M_{\ell b}(\theta_\ell)\}_{\ell=1}^L$, so we have zero uncertainty about the value of the function at the initial evaluation points $\{\theta_\ell\}_{\ell=1}^L$.

[22] See Jones (2001) for an overview of response surface methods for optimization. Convergence results for response surface methods are provided by Bull (2011).

[23] These expressions depend crucially on whether the parameters μ, σ^2, and $\boldsymbol{\beta}$ are consider fixed or random. This is handled somewhat inconsistently in the literature. For instance, Jones et al. (1998) estimate μ, σ^2, and $\boldsymbol{\beta}$ using an empirical Bayes framework, but for analytic convenience they consider only the estimator $\hat{\mu}$ to be random in the calculation of the predictive variance (see Jones et al. (1998) equation (9)). To simplify the discussion, here we treat μ, σ^2, and $\boldsymbol{\beta}$ as fixed, which accounts for the differences in our expressions and those in Jones et al. (1998).

Second, the Gaussian process regression model provides a simple closed-form expression for the predictive variance at new values of θ, which is important in a response surface method to model our uncertainty of the function value at unevaluated points.

To balance the tradeoff between local and global search, Jones et al. (1998) suggest the use of the *expected improvement* criterion for choosing new points. Intuitively, the expected improvement criterion balances the need to search locally around previously obtained minima with the need to search globally at areas of Θ where the predictive variance is high. In particular, if $\hat{\varphi}^{M*}_{\ell b,k}$ is the best value of $\hat{\varphi}^{M}_{\ell b}(\theta)$ obtained up to iteration k, then the expected improvement at a point $\theta \in \Theta$ is given by:

$$E[|\hat{\varphi}^{M*}_{\ell b,k} - \hat{\hat{\varphi}}^{M}_{\ell b}(\theta)|_+] = (\hat{\varphi}^{M*}_{\ell b,k} - \hat{\hat{\varphi}}^{M}_{\ell b}(\theta)) \cdot \Phi\left(\frac{\hat{\varphi}^{M*}_{\ell b,k} - \hat{\hat{\varphi}}^{M}_{\ell b}(\theta)}{\hat{\hat{\sigma}}_{\ell b}(\theta)}\right) + \hat{\hat{\sigma}}_{\ell b}(\theta) \cdot \phi\left(\frac{\hat{\varphi}^{M*}_{\ell b,k} - \hat{\hat{\varphi}}^{M}_{\ell b}(\theta)}{\hat{\hat{\sigma}}_{\ell b}(\theta)}\right),$$

where the expectation is taken with respect to the predictive distribution. At each iteration of the algorithm we choose an evaluation point θ_{L+1} to maximize expected improvement. The middle problem is then evaluated at the new point, the pair $(\theta_{L+1}, \hat{\varphi}^{M}_{\ell b}(\theta_{L+1}))$ is added to the cache $\{(\theta_\ell, \hat{\varphi}^{M}_{\ell b}(\theta_\ell)\}_{\ell=1}^{L}$, the hyperparameters and predictive mean and variances of the Gaussian process regression model are updated, and the procedure repeats. Termination of the algorithm occurs when the expected improvement is below a user-specified threshold.

More details of the implementation are discussed in Section S.4 of the Supplemental Material. We revisit the outer problem in Section 5 where we demonstrate its performance in a practical application.

4 Consistency and Inference

4.1 Consistency

In this section we prove consistency of the plug-in estimator $\hat{\varphi}_{\ell b}$ for the lower bound $\varphi_{\ell b}$ presented in Theorem 2.1. Our objective is to present a consistency result that holds under relatively low-level conditions, which necessarily requires some tradeoffs. In particular, we are able to avoid high-level assumptions on the local problem value functions by restricting the multipliers $\boldsymbol{\lambda}$ to belong to a bounded set $\Lambda \subset \mathbb{R}^{d_m}$. This allows us to obtain uniform (over $\Theta \times \Lambda$) consistency of the average local problem, which in turn is sufficient for consistency of the lower bound. Inspecting Theorem 2.1, restricting $\boldsymbol{\lambda} \in \Lambda$ too much can result in an outer approximation to the identified set Φ^*, although the approximation error can be controlled by the choice of Λ. With a bounded set of multipliers, the remaining assumptions required for uniform consistency are quite weak, which is an advantage of the approach. It may be possible to obtain consistency results without requiring uniform convergence of average local problems, although our own efforts suggest this cannot be done

without substantial additional restrictions on the counterfactual functional, the moment functions and the support restrictions. We leave the extension for future research.

The following assumption is slightly stronger than necessary for consistency, but is also be used to demonstrate the validity of our inference procedure in the next section.

Assumption 4.1. *The function $\varphi_{\ell b}^L(\boldsymbol{y}_i, \boldsymbol{z}_i, \theta, \boldsymbol{\lambda})$ and the parameter space $(\mathcal{P}, \Theta)$ satisfy the following:*

(i) $\{(\boldsymbol{Y}_i, \boldsymbol{Z}_i) : 1 \leq i \leq n\}$ *are i.i.d. under some* $P \in \mathcal{P}$ *on* $(\mathcal{Y} \times \mathcal{Z}, \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}))$.

(ii) $|\varphi_{\ell b}^L(\boldsymbol{y}, \boldsymbol{z}, \theta, \boldsymbol{\lambda})| \leq \overline{M}(\boldsymbol{y}, \boldsymbol{z})$, $\forall (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y} \times \mathcal{Z}$, $\forall \theta \in \Theta$, $\forall \boldsymbol{\lambda} \in \Lambda$ *for a bounded subset* $\Lambda \subset \mathbb{R}^{d_m}$, *for some measurable envelope function* $\overline{M} : \mathcal{Y} \times \mathcal{Z} \to [0, \infty)$ *satisfying* $\mathbb{E}_P[\overline{M}(\boldsymbol{Y}_i, \boldsymbol{Z}_i)^{2+\delta}] \leq C$ *for some* $C < \infty$ *and* $\delta > 0$.

(iii) The class of functions $\boldsymbol{\Phi}_{\ell b}^L := \left\{\varphi_{\ell b}^L(\,\cdot\,, \theta, \boldsymbol{\lambda}) : (\theta, \boldsymbol{\lambda}) \in \Theta \times \Lambda\right\}$ *is pointwise measurable and satisfies Dudley's entropy condition for the envelope* $\overline{M}$.

Assumption 4.1 restricts attention to a setting with i.i.d. data, although the assumption can be relaxed with some modification to the results that follow. Assumption 4.1 imposes that there exists a possibly data-dependent envelope function $\overline{M}$ for the class of functions $\boldsymbol{\Phi}_{\ell b}^L$ that satisfies a standard moment condition. The envelope function is independent of the pair $(\theta, \boldsymbol{\lambda})$ and, as discussed above, the vector $\boldsymbol{\lambda}$ is restricted to lie in a bounded subset $\Lambda \subset \mathbb{R}^{d_m}$. Finally, Assumption 4.1(iii) is a standard condition for uniform consistency results, imposing a weak measurability condition and restricting the complexity of the class of local functions $\boldsymbol{\Phi}_{\ell b}$. Both pointwise measurability and Dudley's entropy condition are defined in Section S.1 of the Supplemental Material. In Section S.5 of the Supplemental Material we provide some results that can be used to verify the entropy condition using primitive assumptions on the counterfactual functional and the moment functions.

The following result demonstrates consistency of the sample analog bounds under Assumption 4.1. Here and in the following section $\Pr_P$ denotes the $n-$fold product measure.

Theorem 4.1. *Suppose Assumption 4.1 holds. Then:*

$$\limsup_{M \to \infty} \limsup_{n \to \infty} Pr_P \left(\left| \inf_{\theta \in \Theta} \sup_{\boldsymbol{\lambda} \in \Lambda} \mathbb{E}_P[\varphi_{\ell b}^L(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \inf_{\theta \in \Theta} \sup_{\boldsymbol{\lambda} \in \Lambda} \hat{\varphi}_{\ell b}^L(\theta, \boldsymbol{\lambda}) \right| > \frac{M}{\sqrt{n}} \right) = 0.$$

Theorem 4.1 follows from Lemma S.2.6 in the Supplemental Material, which proves consistency of $\hat{\varphi}_{\ell b}^L(\theta, \boldsymbol{\lambda})$ for $\mathbb{E}_P[\varphi_{\ell b}^L(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})]$ uniformly over $(\theta, \boldsymbol{\lambda}) \in \Theta \times \boldsymbol{\Lambda}$. The result depends only on Assumption 4.1, and is technically independent of the assumptions in Section 2. In addition to consistency of the plug-in estimator, Theorem 4.1 implies a rate of convergence of $n^{-1/2}$, which is important for the method of inference developed in the next section.

4.2 Inference

In this section we present a feasible but potentially conservative method of confidence set construction. As in the previous section, our inference procedure is designed to avoid high-level assumptions on the local problem value functions. We also focus on designing a procedure that is computationally tractable.

Our aim is to construct an interval $CS_n(1-\alpha) := [CS_n^{\ell b}(1-\alpha), CS_n^{ub}(1-\alpha)]$ that contains the closed convex hull of the identified set $\overline{\text{conv}}(\Phi^*) = [\varphi_{\ell b}, \varphi_{ub}]$ asymptotically with probability at least $1-\alpha$ for some user-specified $\alpha \in (0,1)$:

$$\liminf_{n\to\infty} \Pr_P\left([\varphi_{\ell b}, \varphi_{ub}] \subset CS_n(1-\alpha)\right) \geq 1-\alpha. \tag{4.1}$$

To construct an interval $CS_n(1-\alpha)$ satisfying (4.1), it suffices to construct $CS_n^{\ell b}(1-\alpha)$ and $CS_n^{ub}(1-\alpha)$ to satisfy $\Pr_P(CS_n^{\ell b}(1-\alpha) \geq \varphi_{\ell b}) \leq \alpha/2$ and $\Pr_P(CS_n^{ub}(1-\alpha) \leq \varphi_{ub}) \leq \alpha/2$. We focus on the construction of $CS_n^{\ell b}(1-\alpha)$, as the construction of $CS_n^{ub}(1-\alpha)$ is similar. The value $CS_n^{\ell b}(1-\alpha)$ can be constructed by exploiting the duality between hypothesis testing and confidence set construction. In particular, define:

$$\mathcal{P}_\tau := \left\{ P \in \mathcal{P} : \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P\left[\varphi_{\ell b}^L(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})\right] \leq \tau \right\}, \tag{4.2}$$

and consider the problem of testing $H_0 : P \in \mathcal{P}_\tau$ versus $H_1 : P \in \mathcal{P} \setminus \mathcal{P}_\tau$, where $\mathcal{P}$ is the set of data generating processes satisfying Assumption 4.1 and Assumption 4.2 below. We construct a testing procedure for this null and alternative that controls size asymptotically at level $\alpha/2$, and then set $CS_n^{\ell b}(1-\alpha)$ equal to the smallest value of τ for which we fail to reject the null.

Remark 4.1. *The testing problem has some conceptual similarities to testing problems in the literature on moment inequalities.*

In the absence of the infimum in (4.2), if Λ contains a finite number of elements then the problem $H_0 : P \in \mathcal{P}_\tau$ versus $H_1 : P \in \mathcal{P} \setminus \mathcal{P}_\tau$ is equivalent to testing whether a vector $\theta \in \Theta$ satisfies a finite set of moment inequalities (e.g. Andrews and Soares (2010)). When Λ contains an uncountable set the problem is similar to the testing problem considered in Andrews and Shi (2017).

In the presence of the infimum in (4.2), if Λ contains a finite number of elements then the problem is equivalent to testing whether there exists a vector $\theta \in \Theta$ that satisfies a finite set of moment inequalities. This is the specification-testing environment considered by Bugni et al. (2015), which is also closely related to the problem of subvector inference (see Bugni et al. (2017), Belloni et al. (2019), Kaido et al. (2019)). When Λ contains an uncountable set the problem is similar to the specification testing problem in Marcoux et al. (2023), which is the connection we pursue.

Our test statistic allows a user-specified, data-dependent scale parameter $\hat{\varsigma}_n(\theta)$ that estimates a true scale parameter $\varsigma_P(\theta)$, and satisfies the following assumption.

Assumption 4.2. *(i) For any $\varepsilon > 0$ and any $P \in \mathcal{P}$:*

$$\limsup_{n\to\infty} Pr_P\left(\sup_{\theta\in\Theta}\left|\frac{\varsigma_P(\theta)}{\hat{\varsigma}_n(\theta)} - 1\right| > \varepsilon\right) = 0.$$

(ii) For every $P \in \mathcal{P}$, $\varsigma_P(\theta) \in [\underline{\delta_\varsigma}, \overline{\delta_\varsigma}]$ for some $\underline{\delta_\varsigma} > 0$ and $\overline{\delta_\varsigma} < \infty$, for all $\theta \in \Theta$.

Now consider the function:

$$T_n(\theta) := \sup_{\boldsymbol{\lambda}\in\Lambda} \frac{\sqrt{n}(\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) - \tau)}{\hat{\varsigma}_n(\theta)}, \tag{4.3}$$

and for $\epsilon_n = o(n^{-1/2})$ let $\{\hat{\theta}_n\}_{n=1}^{\infty}$ be any sequence satisfying:

$$T_n(\hat{\theta}_n) \leq \inf_{\theta\in\Theta} T_n(\theta) + \epsilon_n. \tag{4.4}$$

The value $T_n(\hat{\theta}_n)$ serves as our test statistic. Intuitively, when ϵ_n is small, a large value of the test statistic $T_n(\hat{\theta}_n)$ is evidence against the null and in favor of the alternative. Our test function ϕ_n then compares $T_n(\hat{\theta}_n)$ with an appropriately constructed critical value.

As a first pass the critical value might be constructed by approximating the distribution $T_n(\hat{\theta}_n)$ under the null with a resampling procedure that repeatedly computes a version of $T_n(\hat{\theta}_n)$. However, repeatedly re-evaluating the test statistic is computationally demanding for most realistic settings. In particular, although it is typically straightforward to compute the supremum over $\boldsymbol{\lambda} \in \Lambda$ in (4.3) for a fixed value of $\theta \in \Theta$, it is difficult to compute the infimum over $\theta \in \Theta$ in (4.4). To alleviate the computational burden associated with this approach, our proposed procedure reuses the approximate minimizer $\hat{\theta}_n$ when constructing the critical value with the bootstrap. This avoids repeatedly solving an expensive optimization problem, but also introduces additional theoretical challenges. To understand why, suppose $\hat{\varsigma}_n(\hat{\theta}_n) = 1$, and consider the following reformulation of our test statistic:

$$T_n(\hat{\theta}_n) = \sup_{\boldsymbol{\lambda}\in\Lambda}\Bigg\{\underbrace{\sqrt{n}(\hat{\varphi}^L_{\ell b}(\hat{\theta}_n, \boldsymbol{\lambda}) - \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \hat{\theta}_n, \boldsymbol{\lambda})])}_{\text{Empirical Process}} + \underbrace{\sqrt{n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \hat{\theta}_n, \boldsymbol{\lambda})] - \tau)}_{\text{Recentering Term}}\Bigg\}.$$

The first term in the expansion is an empirical process evaluated at $\hat{\theta}_n$, and the distribution of this process can be approximated under weak assumptions using the bootstrap. The second term in the expansion is a recentering term that depends on the estimated $\hat{\theta}_n$. Due to this dependence, and without stronger assumptions, it is possible the recentering term converges in probability under the null to a value that is above zero. This means the test statistic can be asymptotically "too large," posing a challenge for size control.

To eliminate the problem, we use a sample-splitting method proposed by Marcoux et al. (2023) in the context of specification testing for moment inequality models. To appreciate the approach,

first note that Lemma S.2.5 in the Supplemental Material shows, under Assumption 4.1 and 4.2, for any $P \in \mathcal{P}_\tau$ we have:

$$\max\left\{\sup_{\boldsymbol{\lambda}\in\Lambda}\sqrt{n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\hat{\theta}_n,\boldsymbol{\lambda})]-\tau),0\right\}=O_P(1).$$

That is, without introducing any additional assumptions, the minimizer $\hat{\theta}_n$ is of sufficient quality that the recentering term is stochastically bounded. As a result, for any sequence $q_n = o(n)$ we have:

$$\max\left\{\sup_{\boldsymbol{\lambda}\in\Lambda}\sqrt{q_n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\hat{\theta}_n,\boldsymbol{\lambda})]-\tau),0\right\}=o_P(1).$$

In other words, after multiplication by any sequence $q_n/n = o(1)$ the recentering term either diverges to $-\infty$, or converges in probability to a term bounded above by zero. This prevents the recentering term and test statistic from being "too large" asymptotically, and motivates the following modified test statistic:

$$T_{q_n}(\hat{\theta}_n) := \sup_{\boldsymbol{\lambda}\in\Lambda}\frac{\sqrt{q_n}(\hat{\varphi}^L_{\ell b,q_n}(\hat{\theta}_n,\boldsymbol{\lambda})-\tau)}{\hat{\varsigma}_n(\hat{\theta}_n)}, \tag{4.5}$$

where $\hat{\theta}_n$ is computed as before, and where $\hat{\varphi}^L_{\ell b,q_n}(\hat{\theta}_n,\boldsymbol{\lambda})$ is the sample average of the local function $\varphi^L_{\ell b}(\,\cdot\,,\hat{\theta}_n,\boldsymbol{\lambda})$ taken over q_n randomly sampled observations. Expanding the modified test statistic in the case when $\hat{\varsigma}_n(\hat{\theta}_n)=1$ we now obtain:

$$T_{q_n}(\hat{\theta}_n)=\sup_{\boldsymbol{\lambda}\in\Lambda}\Bigg\{\underbrace{\sqrt{q_n}(\hat{\varphi}^L_{\ell b,q_n}(\hat{\theta}_n,\boldsymbol{\lambda})-\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\hat{\theta}_n,\boldsymbol{\lambda})])}_{\text{Empirical Process}}+\underbrace{\sqrt{q_n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\hat{\theta}_n,\boldsymbol{\lambda})]-\tau)}_{\text{Modified Recentering Term}}\Bigg\}.$$

Again, the distribution of the empirical process can be approximated using the bootstrap, but now the modified recentering term tends to zero in probability under the null. Furthermore, it can be shown the modified recentering term diverges for any fixed alternative $P \in \mathcal{P}\setminus\mathcal{P}_\tau$ while the empirical process remains stochastically bounded, providing the basis for power against fixed alternatives.

The modified test statistic $T_{q_n}(\hat{\theta}_n)$ requires the full sample to compute $\hat{\theta}_n$, but otherwise uses only a small (relative to n) subsample of the full sample to compute. While the test has favorable asymptotic properties, the power of the test may be improved by making better use of the full sample. Our final proposed test addresses the issue by using multiple test statistics constructed using different subsamples. The final testing procedure is as follows:

Step 1: Compute $\hat{\varphi}^L_{\ell b}(\theta,\boldsymbol{\lambda})$ and $\hat{\varsigma}_n(\theta)$, and find an approximate minimizer $\hat{\theta}_n$ (in the sense of (4.4)) of the function $T_n(\theta)$ from (4.3).

Step 2: Fix q_n and r_n satisfying $q_n = o(n)$ and $r_n \cdot q_n \leq n$, where r_n is a bounded but possible in-

creasing sequence. Divide the sample $\{(\boldsymbol{Y}_i, \boldsymbol{Z}_i)\}_{i=1}^n$ into r_n subsamples $\{(\boldsymbol{Y}_i^{(r)}, \boldsymbol{Z}_i^{(r)})\}_{i=1}^{q_n}$ of size q_n, and compute $T_{q_n}^{(1)}(\hat{\theta}_n), \ldots, T_{q_n}^{(r_n)}(\hat{\theta}_n)$ using each subsample.

Step 3: Fix some large integer B. For $r = 1, \ldots, r_n$ resample B bootstrap samples $\{\{(\boldsymbol{Y}_{i,b}^{(r)}, \boldsymbol{Z}_{i,b}^{(r)})\}_{i=1}^n : b = 1, \ldots, B\}$ i.i.d. with replacement from $\{(\boldsymbol{Y}_i^{(r)}, \boldsymbol{Z}_i^{(r)})\}_{i=1}^{q_n}$ and compute the bootstrap test statistic:

$$T_{q_n,b}^{(r)}(\hat{\theta}_n) := \sup_{\boldsymbol{\lambda} \in \Lambda} \frac{\sqrt{q_n}(\hat{\varphi}_{\ell b,b}^{L(r)}(\hat{\theta}_n, \boldsymbol{\lambda}) - \hat{\varphi}_{\ell b}^{L(r)}(\hat{\theta}_n, \boldsymbol{\lambda}))}{\hat{\varsigma}_n(\hat{\theta}_n)}, \tag{4.6}$$

where:

$$\hat{\varphi}_{\ell b}^{L(r)}(\theta, \boldsymbol{\lambda}) := \frac{1}{q_n} \sum_{i=1}^{q_n} \varphi_{\ell b}^L(\boldsymbol{Y}_i^{(r)}, \boldsymbol{Z}_i^{(r)}, \theta, \boldsymbol{\lambda}), \quad \hat{\varphi}_{\ell b,b}^{L(r)}(\theta, \boldsymbol{\lambda}) := \frac{1}{q_n} \sum_{i=1}^{q_n} \varphi_{\ell b}^L(\boldsymbol{Y}_{i,b}^{(r)}, \boldsymbol{Z}_{i,b}^{(r)}, \theta, \boldsymbol{\lambda}).$$

Step 4: Fix some infinitesimal $\eta > 0$, and for $r = 1, \ldots, r_n$ choose $c_n^{(r)}(1 - \alpha/r_n + \eta)$ as the $1 - \alpha/r_n + \eta$ quantile of the bootstrap distribution of $T_{q_n,b}^{(r)}(\hat{\theta}_n)$.[24]

Step 5: Reject the null hypothesis if $T_{q_n}^{(r)}(\hat{\theta}_n) > c_n^{(r)}(1 - \alpha/r_n + \eta) + \eta$ for any $r = 1, \ldots, r_n$.

Summarizing, our final test function is:

$$\phi_n = \bigvee_{r=1}^{r_n} \mathbb{1}\left\{T_{q_n}^{(r)}(\hat{\theta}_n) > c_n^{(r)}(1 - \alpha/r_n + \eta) + \eta\right\}.$$

The aggregation of the test statistics is akin to a Bonferroni correction, which is known to be conservative.[25] However, in some cases the final test function can offer power improvements over a single test of the null hypothesis using only one of the subsamples. The following theorem summarizes the size control and power properties of our testing procedure proposed above. Here $\mathbb{P}^\sharp$ represents the uncertainty from the resampling procedure, and is defined formally in Section S.1 of the Supplemental Material.

Theorem 4.2. *Suppose Assumptions 4.1 and 4.2 hold. Then for any $P \in \mathcal{P}_\tau$:*

$$\limsup_{n \to \infty} (Pr_P \times \mathbb{P}^\sharp)(\phi_n = 1) \leq \alpha.$$

Furthermore, for any $P \in \mathcal{P} \setminus \mathcal{P}_\tau$:

$$\liminf_{n \to \infty} (Pr_P \times \mathbb{P}^\sharp)(\phi_n = 1) = 1.$$

[24] Here η is the infinitesimal uniformity factor in Andrews and Shi (2017), which is required to avoid certain high level assumptions on the asymptotic distribution of the test statistics. See Andrews and Shi (2017) footnote 20.

[25] Note the dependence structure of the test statistics, as well as the fact that we are testing a common null (rather than distinct nulls) separates the problem from the multiple testing literature. Some alternatives to the Bonferroni correction for aggregating multiple tests of the same null have been studied in Vovk and Wang (2020), and a full study of the power properties of these alternative approaches is a direction for future research.

The proof of Theorem 4.2 shows the proposed test controls size, and has power tending to 1 for fixed alternatives. Our final confidence set constructed based on test inversion inherits the properties of the proposed test, satisfying (4.1) and asymptotically excluding any point outside of the interval $[\varphi_{\ell b}, \varphi_{ub}]$. Computing our modified test statistic in (4.5) requires solving an optimization problem over $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ for each bootstrap sample, which can be solved using the same methods proposed in Section 3.2. Computing the bootstrap test statistic in (4.6) is slightly more complicated, since it is generally not a concave problem, and instead involves the difference of two concave functions. Maximizing a difference of two concave functions is the same as minimizing the difference of two convex functions, for which a number of efficient and globally convergent algorithms are available when the subgradients of the functions are known (see Le Thi and Pham Dinh (2018)).[26] Stochastic analogs of these algorithms have also been recently developed and studied by Le Thi et al. (2022). In the end, similar techniques described in Section 3.2 can be used to solve the problem in (4.6) for each bootstrap sample.

5 Application: Revisiting Airline Competition

In this section we apply the method to the airline entry data from Ciliberto and Tamer (2009), investigating the impact of parametric distributional assumptions on their results. As described in Ciliberto and Tamer (2009), the data are taken from the second quarter of the 2001 Airline Origin and Destination Survey. Each observation represents a market, defined as a trip between two airports.[27] The data include 2742 markets. We observe whether or not the market is served by American Airlines (AA), Delta Air Lines (DL), United Airlines (UA), and Southwest Airlines (WN), as well as whether the market is served by a medium airline (MA) (either America West, Continental, Northwest or USAir), or a low cost carrier (LC). As in Ciliberto and Tamer (2009) we suppose the airlines decide whether to service a route between two airports according to a binary game of complete information with pure strategy Nash equilibria. In particular, $Y_{ik} \in \{0, 1\}$ represents the entry decision of airline $k \in \{AA, DL, UA, WN, MA, LC\}$ in market i, and satisfies:

$$Y_{ik} = \mathbb{1}\left\{ \boldsymbol{Z}_{ik}^{\top}\beta + \delta \sum_{k' \neq k} Y_{ik'} \geq U_{ik} \right\},$$

where $\boldsymbol{Z}_{ik}$ is a vector of covariates, $\boldsymbol{Y}_{i(-k)} \in \{0, 1\}^5$ is a vector whose components indicate the entry decisions of each of the other airlines, and U_{ik} is a market- and player-specific payoff-relevant latent variable. In our specification, $\boldsymbol{Z}_{ik} = (1, MS_i, MP_{ik}, WA_i)$ is a 4×1 vector. The covariate

[26] Briefly, in the simplest algorithms, the difference of convex functions $f(\boldsymbol{x}) - h(\boldsymbol{x})$ can be minimized by repeatedly (i) updating the subgradient $\boldsymbol{g}_h^k \in \partial h(\boldsymbol{x}^k)$, and (ii) updating the iterates $\boldsymbol{x}^{k+1} = \arg\min\{f(\boldsymbol{x}) - \langle \boldsymbol{g}_h^k, \boldsymbol{x} \rangle\}$. Note step (ii) is a convex minimization, where $h(\boldsymbol{x})$ has been replaced in the difference $f(\boldsymbol{x}) - h(\boldsymbol{x})$ with its linear approximation.

[27] Ignoring transfers and the direction of the flight.

MS_i is a standardized variable representing the market size, defined as the geometric mean of the populations at the two market endpoints. The variable MP_{ik} is a standardized variable representing airline k's presence in market i, which is the average proportion of markets served by airline k in the origin and destination airports. Both MS_i and MP_i are variables with infinite support. Finally, the variable WA_i is a binary variable equal to 1 if market i was affected the Wright Amendment legislation.

The Wright Amendment was passed by Congress in 1979, but was fully repealed in 2014. The goal of the legislation was to stimulate growth out of the Dallas/Fort Worth airport by restricting the markets served out of Dallas Love. Only markets located in Texas, Louisiana, Arkansas, Oklahoma, New Mexico, Alabama, Kansas, and Mississippi could be served out of Dallas Love under the Wright Amendment. In the data, 93 markets are served out of Dallas Love, and 81 of the markets are affected by the Wright Amendment. Following Ciliberto and Tamer (2009), we analyze the impact of the repeal of the Wright Amendment on the entry of various airlines into markets serving routes from Dallas Love. In particular, our counterfactual modifies $\boldsymbol{Z}_{ik}$ to be $\check{\boldsymbol{Z}}_{ik}$, where $\check{\boldsymbol{Z}}_{ik} = (1, MS_i, MP_{ik}, 0)$; that is, where the Wright Amendment dummy variable has been set to zero for all markets. Our counterfactual outcome variable is then given by:

$$Y_{ik}^{\star} = \mathbb{1}\left\{ \check{\boldsymbol{Z}}_{ik}^{\top}\beta + \delta \sum_{k' \neq k} Y_{ik'}^{\star} \geq U_{ik} \right\},$$

for all $k \in \{AA, DL, UA, WN, MA, LC\}$. Our counterfactual parameter of interest is the counterfactual conditional entry probabilities $\mathbb{P}(Y_{ik}^{\star} = 1 \mid WA_i = 1)$ for each player $k \in \{AA, UA\}$.[28] Our objective is to study the sensitivity of the counterfactual entry probabilities to different assumptions on the number of strategically interacting players and to parametric distributional assumptions on the latent variables. We consider two scenarios: in the first we consider the full $6-$player game, and in the second we consider only a two-player game between AA and UA. In the two-player game, only AA and UA interact strategically, and all other airlines make a simple binary choice of whether to enter without considering the actions of the other players.[29] We also consider two baseline distributions for the latent variables. For the first baseline distribution, we suppose $U'_{ik} \sim N(0,1)$ for $k \in \{AA, DL, UA, WN, MA, LC\}$, but we do not impose any structure on the dependence between U'_{ik} and $U'_{ik'}$ for $k \neq k'$. For the second baseline distribution, we impose $\boldsymbol{U}'_i \sim N(\boldsymbol{0}, \boldsymbol{I})$, forcing the payoff relevant latent variables to be independent across all players. For both configurations, we assume the true vector of latent variables $\boldsymbol{U}_i$ lie in a $1-$Wasserstein neighborhood of radius ρ

[28] This parameter can be bounded using our framework by setting $\varphi(\boldsymbol{Y}_i^{\star}, \boldsymbol{Z}_i, \boldsymbol{U}_i) = Y_{ik}^{\star} \cdot \mathbb{1}\{WA_i = 1\}$, and then by rescaling the sample lower and upper bounds by $n/\sum_{i=1}^{n} \mathbb{1}\{WA_i = 1\}$. Note this objective is slightly different from the objective in Ciliberto and Tamer (2009), who focus on counterfactual entry probabilities in markets in or out of Dallas Love (93 markets), which is a superset of the set of markets affected by the Wright Amendment (81 markets).

[29] That is, only UA's action affects the payoff of AA and vice versa. For all other players the actions of the other players do no enter their payoff functions.

	Six Players, Dependent U'_{ik}					Six Players, Independent U'_{ik}			
	AA		UA			AA		UA	
ρ	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	ρ	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$
0	$+\infty$	$-\infty$	$+\infty$	$-\infty$	0.83	$+\infty$	$-\infty$	$+\infty$	$-\infty$
0.01	0.00	1.00	0.00	1.00	0.84	0.04	0.40	0.09	1.00
0.02	0.00	1.00	0.00	1.00	0.85	0.04	0.96	0.00	1.00
0.03	0.00	1.00	0.00	1.00	0.86	0.02	1.00	0.00	1.00
0.04	0.00	1.00	0.00	1.00	0.87	0.00	1.00	0.00	1.00
0.05	0.00	1.00	0.00	1.00	0.88	0.00	1.00	0.00	1.00

	Two Players, Dependent U'_{ik}					Two Players, Independent U'_{ik}			
	AA		UA			AA		UA	
ρ	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	ρ	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$	$\hat{\varphi}_{\ell b}$	$\hat{\varphi}_{ub}$
0	$+\infty$	$-\infty$	$+\infty$	$-\infty$	0.25	$+\infty$	$-\infty$	$+\infty$	$-\infty$
0.01	0.00	1.00	0.00	1.00	0.26	0.04	0.63	0.00	0.46
0.02	0.00	1.00	0.00	1.00	0.27	0.01	0.68	0.00	0.66
0.03	0.00	1.00	0.00	1.00	0.28	0.00	1.00	0.00	0.69
0.04	0.00	1.00	0.00	1.00	0.29	0.00	1.00	0.00	0.93
0.05	0.00	1.00	0.00	1.00	0.30	0.00	1.00	0.00	0.97

Table 1: Bounds on the counterfactual entry probabilities of AA and UA for various values of ρ after repealing the Wright Amendment. The bounds are constructed under various assumptions on number of strategically interacting players, as well as under different assumptions on the dependence structure of the baseline distribution for the latent variables.

around the baseline distribution, and we study the effects on the counterfactual entry probabilities as we vary the value of ρ. We also impose sign restrictions on the structural parameters, restricting $\delta \leq 0$, and restricting the coefficient on MS_i and MP_{ik} to be positive and the coefficient on WA_i to be negative. Importantly, we do not impose any assumptions on how the airlines select among multiple equilibria.

The bounds on counterfactual entry probabilities for various values of ρ are displayed in Table 1. The reported bounds were constructed using the Adam algorithm for the middle problem, and focus on ranges of ρ which lead to nonempty bounds. The results without imposing independence of the player-specific payoff shocks $U'_{ik} \sim N(0,1)$ in the baseline distribution are labelled "Dependent U'_{ik}" in Table 1. For both the six-player and two-player games the bounds are empty at $\rho = 0$ and uninformative for $\rho \geq 0.01$. Empty bounds at $\rho = 0$ does not mean the assumption $U_{ik} \sim N(0,1)$ for all k is rejected, since our latent variable matching procedure relies on a finite number of draws $U'_{ik} \sim N(0,1)$, and so is subject to sampling uncertainty.[30] The uninformative bounds for $\rho \geq 0.01$ suggest that, when the payoff-relevant latent variables are allowed to be arbitrarily correlated across players, it is impossible to predict the effects of repealing the Wright Amendment on the

[30]For instance, consider the 1−Wasserstein distance between two empirical distributions based on two samples drawn from a standard normal. Since the two samples are both from a standard normal, asymptotically the 1−Wasserstein distance between the empirical distributions tends to zero. However, for $n = 1000$ the expected 1−Wasserstein distance between the two empirical distributions is about 0.0571.

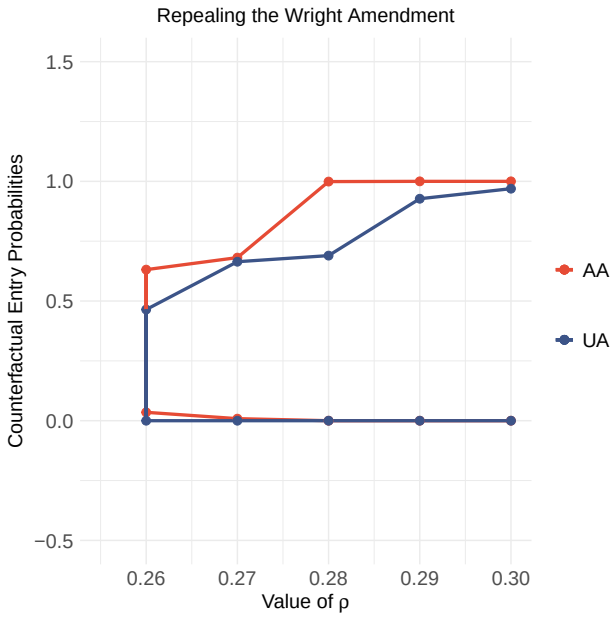
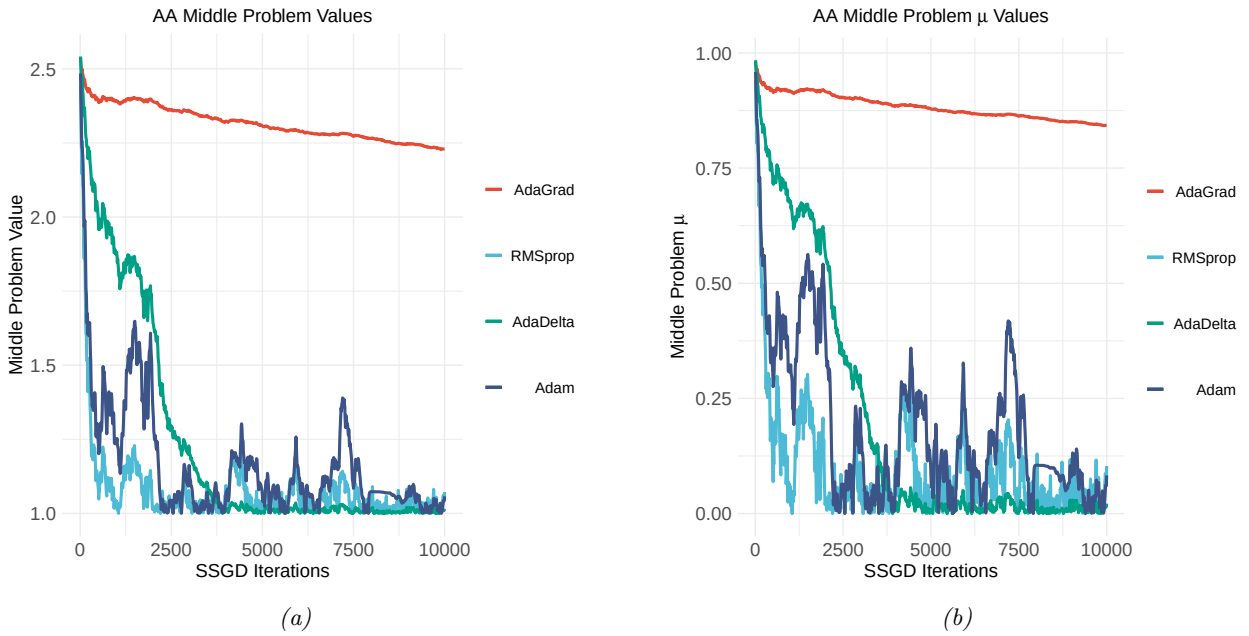


Figure 4: The counterfactual entry probabilities for AA and UA in the two-player game with independent baseline latent variables U'_{ik} for various values of ρ, the upper bound on the $1-$Wasserstein neighborhood. For $\rho < 0.26$ the bounds are empty, and for the $\rho > 0.30$ the bounds are $[0, 1]$. For intermediate values of ρ, the bounds gradually widen as the $1-$Wasserstein neighborhood becomes larger.

entry probabilities of AA and UA.[31]

Further investigation of the two-player game reveals the result is also driven by the fact that (i) the coefficient on the Wright Amendment variable that obtains the upper bound is negative and large (in absolute value) relative to the variance of U'_{ik}, and (ii) there are no markets affected by the Wright Amendment where both AA and UA enter in the data. The first fact implies that when the Wright Amendment is repealed every player receives a large and positive payoff shock in the markets affected by the Wright Amendment. Ceteris paribus, this greatly increases the probability of entering markets affected by the Wright Amendment in the counterfactual, an effect which dominates in the upper bound. For the lower bound, since the coefficient on the Wright Amendment is constrained to be negative, the Dallas Love markets affected by the legislation that were serviced by AA and UA in the data will continue to be serviced by AA and UA in the counterfactual. However, there are no Dallas Love markets affected by the Wright Amendment that were serviced by both AA and UA in the data, allowing for the possibility of a zero lower bound.

[31] Note the results do not contradict the results in Ciliberto and Tamer (2009), since both our specification and our counterfactual parameter is different. Furthermore, although Ciliberto and Tamer (2009) report marginal effects, their specification of the model using moment inequalities does not allow them to compute the true marginal effect in the presence of multiple equilibria. Instead their reported marginal effects are based on the maximum change in the average upper bound of observing a carrier in any possible market structure. See the discussion on pp. 1820-1821 of Ciliberto and Tamer (2009). Our ability to bound true marginal effects in the presence of multiple equilibria is another advantage of our approach.

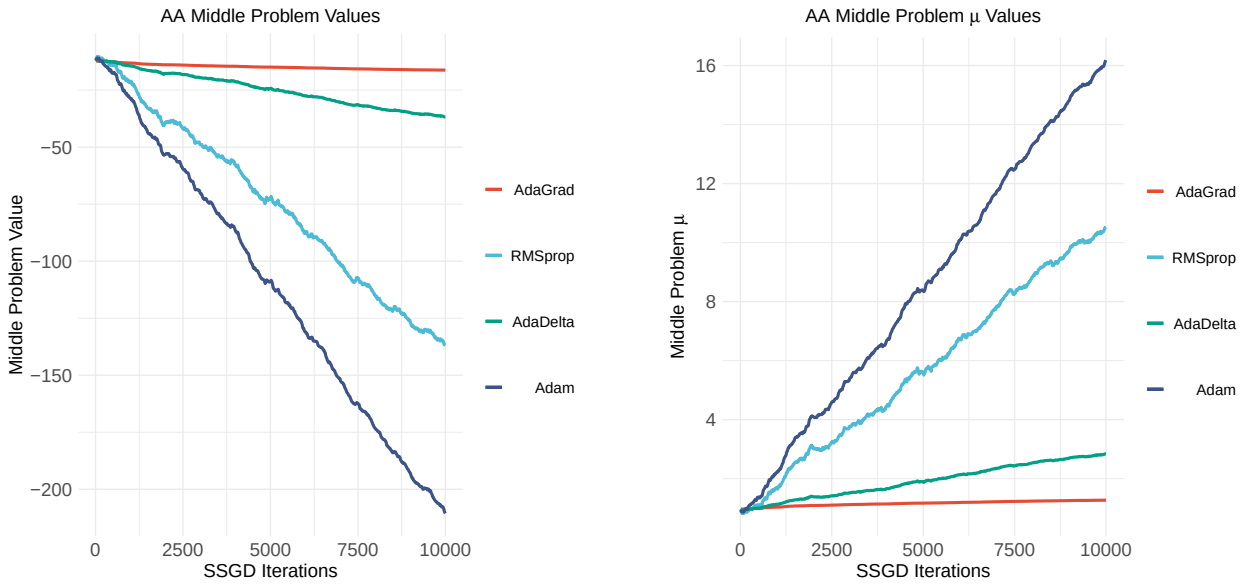


(a) *(b)*

Figure 5: Example results from benchmarking the four SSGD algorithms introduced in Section 3.2 at a value of θ that is inside the identified set when computing the upper bound on counterfactual entry probabilities in the six-player game example without imposing independence. Figure 5(a) shows the value of the middle problem over the first 10^4 iterations, all of which are converging to the upper bound of 1. Figure 5(b) shows the corresponding values of the multiplier μ, all of which are converging to near 0.

The results are different for the specifications where we draw $\boldsymbol{U}_i' \sim N(\mathbf{0}, \boldsymbol{I})$, which forces the payoff-relevant latent variables to be independent across players in the baseline distribution. These results are labelled "Independent U_{ik}'" in Table 1. Here we find a much higher threshold of ρ is required before the bounds become nonempty, consistent with the fact the baseline distribution with independent payoff shocks imposes stronger restrictions. To illustrate the deviations from the baseline distribution, Figure 1 in Section S.5 of the Supplemental Material shows quantile-quantile plots comparing the quantiles of U_{ik}, obtained from the solution to the local problems, to the quantiles of a standard normal for all six players when maximizing the counterfactual entry probability of AA with $\rho = 0.84$. The plots show the success of the matching procedure, with nearly all players' latent variable distributions close to the baseline distribution. An exception is the low cost carriers, which show substantial deviations from the standard normal at the upper quantiles. Unlike the case of dependent payoff shocks the results are informative for a range of ρ, and an illustration of the bounds for the two-player case is provided in Figure 4. After repealing the Wright Amendment the smallest nonempty bounds on the counterfactual entry probabilities in markets affected by the Wright Amendment in the two-player game are $[0.04, 0.63]$ for AA and $[0.00, 0.46]$ for UA.

We also use the application to study the practical performance of our proposed algorithm. Figures 5 and 6 show example results from our benchmarking of the four SSGD algorithms from



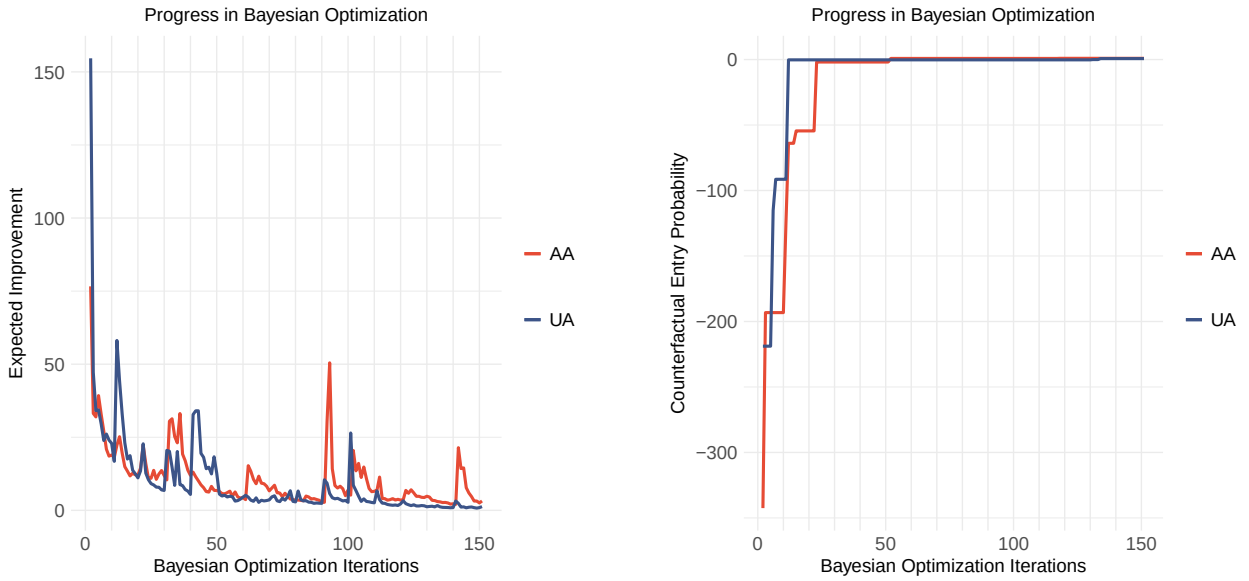
(a) *(b)*

Figure 6: Example results from benchmarking the four SSGD algorithms introduced in Section 3.2 at a value of θ that is outside the identified set when computing the upper bound on counterfactual entry probabilities in the six-player game example without imposing independence. Figure 6(a) shows the value of the middle problem over the first 10^4 iterations, all of which are diverging to $-\infty$. Figure 6(b) shows the corresponding values of the multiplier μ, almost all of which are diverging to $+\infty$.

Section 3.2. In both figures, we are maximizing the counterfactual entry probability of AA without imposing independence of U'_{ik} across players at a value of ρ where the upper bound is 1. For the sake of illustration, we set the initial value of μ to be 1.[32] Figure 5 shows the values of the middle problem and the value of μ during the first 10^4 iterations of SSGD at a vector θ that is feasible and obtains the upper bound of 1. Figure 5(a) shows that three of four SSGD algorithms converge quickly to the upper bound of 1, and Figure 5(b) shows μ converging to a value near zero. The exception is AdaGrad, which exhibits slow progress due to its quickly decaying step size. In contrast, Figure 6 shows the values of the middle problem and the value of μ at a vector θ that is infeasible, with an upper bound tending to $-\infty$. Figure 6(a) shows most of the SSGD algorithms diverge quickly to $-\infty$, and Figure 6(b) shows μ diverging quickly to $+\infty$. Again the exception is AdaGrad. Since RMSprop, AdaDelta and Adam have similar performance for $\theta \in \Theta^*$, we use Adam to construct all bounds because of its ability to quickly diverge for values $\theta \notin \Theta^*$.

Finally, we use the application to investigate the performance of the outer optimization problem over $\theta \in \Theta$. Figure 7 shows the performance of the Bayesian optimization implementation of the outer problem in the six-player example without independent baseline latent variables when maximizing both AA's and UA's counterfactual entry probability for $\rho = 0.84$. Bayesian optimization

[32] In practice the initial value is determined after a preliminary grid search, as described in Section S.4 of the Supplemental Material.



(a) *(b)*

Figure 7: An example of progress during Bayesian optimization when computing the upper bound on counterfactual probabilities in the six-player game example without imposing independence. Figure 7(a) shows the drop in expected improvement over 150 iterations of Bayesian optimization, and Figure 7(b) shows the corresponding improvement in the objective function.

begins after an initial evaluation of 51 grid points over Θ. Figure 7(a) shows that after the initial evaluation the expected improvement from further exploration can be quite high. Furthermore, before starting Bayesian optimization, Figure 7(b) shows that the value of AA's and UA's upper bound lies far below zero. Bayesian optimization is able to quickly find promising evaluation points, and within an additional 51 iterations makes rapid progress beyond the initial grid search. After less than 150 iterations, both AA's and UA's upper bound problem converges to 1. Note the expected improvement in Figure 7(a) does not decline monotonically. Every 10 iterations we re-estimate the hyperparameters for the Gaussian process regression model, which can cause expected improvement to jump before declining again.

6 Conclusion

This paper studies the problem of constructing bounds on scalar counterfactual parameters in partially identified structural models while relaxing parametric distributional assumptions on the latent variables. Using concepts from random set theory and convex analysis, we derive the dual form of the counterfactual bounds, and show how to investigate sensitivity to a baseline distribution using the Wasserstein distance. We explore computational issues in detail, and demonstrate that the dual bounds can be estimated by solving a sequence of nested, but relatively simple optimization problems. We then propose a simple inference procedure, and show the performance of the bounds

in an application to airline entry games.

This paper focuses on general issues and general-purpose algorithms, although a number of special cases of the framework admit substantial simplifications. Exploiting additional structure that arises in specific problems is a promising next step, and may lead to further breakthroughs in estimation and inference. This paper is also part of a growing literature that seeks to use advances in optimization and computer science to improve economic models. A number of nontrivial issues are raised when estimating structural models without parametric distributional assumptions, but we believe the cause is worthwhile and we hope this work inspires others to confront these new challenges.

References

Adjaho, C. and Christensen, T. (2022). Externally valid treatment choice. *arXiv preprint arXiv:2205.05561*.

Andrews, D. W. and Shi, X. (2017). Inference based on many conditional moment inequalities. *Journal of econometrics*, 196(2):275–287.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Arellano, M. and Bonhomme, S. (2021). Recovering latent variables by matching. *Journal of the American Statistical Association*, pages 1–14.

Belloni, A., Bugni, F. A., and Chernozhukov, V. (2019). Subvector inference in pi models with many moment inequalities. Technical report, cemmap working paper.

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

Bertsekas, D. (2009). *Convex optimization theory*, volume 1. Athena Scientific.

Blanchet, J., Kang, Y., and Murthy, K. (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.

Bugni, F. A., Canay, I. A., and Shi, X. (2015). Specification tests for partially identified models defined by moment inequalities. *Journal of Econometrics*, 185(1):259–282.

Bugni, F. A., Canay, I. A., and Shi, X. (2017). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8(1):1–38.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10).

Champion, T. and De Pascale, L. (2011). The monge problem in r d.

Chen, X., Hansen, L. P., and Hansen, P. G. (2021). Robust inference for moment condition models without rational expectations. *Journal of Econometrics, forthcoming*.

Chesher, A. and Rosen, A. M. (2021). Counterfactual worlds. *Annals of Economics and Statistics*, (142):311–335.

Christensen, T. and Connault, B. (2023). Counterfactual sensitivity and robustness. *Econometrica*, 91(1):263–298.

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

De Paula, Á. (2020). Econometric models of network formation. *Annual Review of Economics*, 12:775–799.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

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

Fan, Y., Park, H., and Xu, G. (2023). Quantifying distributional model risk in marginal problems via optimal transport. *arXiv preprint arXiv:2307.00779*.

Gao, W. Y., Li, M., and Xu, S. (2022). Logical differencing in dyadic network formation models with nontransferable utilities. *Journal of Econometrics*.

Gu, J., Russell, T., and Stringham, T. (2022). Counterfactual identification and latent space enumeration in discrete outcome models. *Available at SSRN 4188109*.

Gu, J. and Russell, T. M. (2022). Partial identification in nonseparable binary response models with endogenous regressors. *Journal of Econometrics*.

Gualdani, C. (2021). An econometric model of network formation with an application to board interlocks between firms. *Journal of Econometrics*, 224(2):345–370.

Gualdani, C. and Sinha, S. (2020). Partial identification in matching models for the marriage market.

Hansen, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *American Economic Review*, 91(2):60–66.

Henry, M. and Mourifié, I. (2013). Set inference in latent variables models. *The Econometrics Journal*, 16(1):S93–S105.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.

Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Kalouptsidi, M., Kitamura, Y., Lima, L., and Souza-Rodrigues, E. (2021). Counterfactual analysis for structural dynamic discrete choice models. Technical report, Working paper, Harvard University.

Kédagni, D., Li, L., and Mourifié, I. (2020). Discordant relaxations of misspecified models. *arXiv preprint arXiv:2012.11679*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Le Thi, H. A., Huynh, V. N., Dinh, T. P., and Hau Luu, H. P. (2022). Stochastic difference-of-convex-functions algorithms for nonconvex programming. *SIAM Journal on Optimization*, 32(3):2263–2293.

Le Thi, H. A. and Pham Dinh, T. (2018). Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68.

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31.

Lee, W. (2022). *Identification and estimation of dynamic random coefficient models*. PhD thesis.

Li, L. (2021). Identification of structural and counterfactual parameters in a large class of structural econometric models. Working paper.

Maccheroni, F., Marinacci, M., and Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498.

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

Marcoux, M., Russell, T. M., and Wan, Y. (2023). A simple specification test for models with many conditional moment inequalities. *SSRN preprint ssrn.4345300*.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166.

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

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pflug, G. and Wozabal, D. (2007). Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442.

Schennach, S. M. and Starck, V. (2022). *Optimally-transported generalized method of moments*. Cemmap, Centre for Microdata Methods and Practice, The Institute for Fiscal ....

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275.

Sheng, S. (2020). A structural econometric analysis of network formation games through subnetworks. *Econometrica*, 88(5):1829–1858.

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Tebaldi, P., Torgovitsky, A., and Yang, H. (2023). Nonparametric estimates of demand in the california health insurance exchange. *Econometrica*, 91(1):107–146.

Tieleman, T. and Hinton, G. (2012). Rmsprop, coursera: Neural networks for machine learning. *Technical report*.

Villani, C. (2009). *Optimal transport: old and new*. Springer.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Wozabal, D. (2012). A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Supplementary Material for "A Dual Approach to Wasserstein-Robust Counterfactuals"

Jiaying Gu
University of Toronto

Thomas M. Russell
Carleton University

July 21, 2023

S.1 Background Material

This section provides the minimal background required to understand the main assumptions, results and proofs.

S.1.1 Random Set Theory

Definition S.1.1 (Weak Measurability, Random Closed Set)**.** *Let $(\mathcal{X}, \mathfrak{X})$ be a measurable space and let $\mathcal{S}$ be a topological space. Then the multifunction $\mathcal{F} : \mathcal{X} \to \mathcal{S}$ is said to be weakly measurable if:*

$$\{\boldsymbol{x} \in \mathcal{X} : \mathcal{F}(\boldsymbol{x}) \cap G \neq \varnothing\} \in \mathfrak{X},$$

for every open subset G of S. If $\mathcal{F}(\boldsymbol{x})$ is also closed for all $\boldsymbol{x} \in \mathcal{X}$, then $\mathcal{F}(\boldsymbol{x})$ is called a random closed set.

Remark S.1.1. *When the set $\mathcal{F}(\boldsymbol{x})$ is closed for all $\boldsymbol{x} \in \mathcal{X}$, weak measurability is equivalent to Effros measurability. When $\mathcal{S}$ is locally compact and Hausdorff, the open sets in Definition S.1.1 can be replaced with compact sets.*

Definition S.1.2 (Carathéodory Function)**.** *Let $(\mathcal{X}, \mathfrak{X})$ be a measurable space, and let $\mathcal{S}$ and $\mathcal{T}$ be topological spaces equipped with the Borel $\sigma-$algebra. A function $f : \mathcal{X} \times \mathcal{S} \to \mathcal{T}$ is a Carathéodory function if $f(\cdot, \boldsymbol{s}) : \mathcal{X} \to \mathcal{T}$ is measurable for each $\boldsymbol{s} \in \mathcal{S}$, and $f(\boldsymbol{x}, \cdot) : \mathcal{S} \to \mathcal{T}$ is continuous for each $\boldsymbol{x} \in \mathcal{X}$.*

All errors are our own.

S.1.2 Convex Analysis

Definition S.1.3 (Affine Hull, Relative Interior)**.** *Given a set $A \subset \mathbb{R}^d$, the affine hull of A, denoted by aff(A), is given by:*

$$\mathit{aff}(A) := \left\{ \sum_{k=1}^{K} \alpha_k \boldsymbol{x}_k \;\middle|\; K > 0,\ \boldsymbol{x}_k \in A,\ \alpha_k \in \mathbb{R}, \sum_{k=1}^{K} \alpha_k = 1 \right\}.$$

Furthermore, the relative interior of A, denoted by ri(A), is the set consisting of the interior points of A when it is considered as a subset of aff(S):

$$\mathit{ri}(S) = \{\boldsymbol{x} \in \mathbb{R}^d : \exists \varepsilon > 0 \ s.t.\ B(\boldsymbol{x}, \varepsilon) \cap \mathit{aff}(A) \subseteq A\},$$

where $B(\boldsymbol{x}, \varepsilon)$ is the open ball of radius $\varepsilon > 0$ around $\boldsymbol{x} \in \mathbb{R}^d$.

Definition S.1.4 (Proper Function)**.** *A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is proper if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $f(\boldsymbol{x}) < +\infty$ for some $\boldsymbol{x} \in \mathbb{R}^d$.*

Definition S.1.5 (Convex Conjugate)**.** *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper convex function. Then the convex conjugate of f, denoted by $f^* : \mathbb{R}^d \to \overline{\mathbb{R}}$, is given by:*

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} \{\langle \boldsymbol{x}, \boldsymbol{x}^* \rangle - f(\boldsymbol{x})\}.$$

Definition S.1.6 (Subgradient, Subdifferential)**.** *Given a proper, convex function $f : \mathbb{R}^d \to \mathbb{R}$, a subgradient of f at $\boldsymbol{x} \in \mathbb{R}^d$ is any vector $g \in \mathbb{R}^d$ satisfying:*

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top (\boldsymbol{z} - \boldsymbol{x}), \qquad \forall \boldsymbol{z} \in \mathbb{R}^d.$$

The subdifferential of f at $\boldsymbol{x} \in \mathbb{R}^d$, denoted by $\partial f(\boldsymbol{x})$, is the set of all subgradients of f at $\boldsymbol{x} \in \mathbb{R}^d$. We say that f is subdifferentiable at $\boldsymbol{x}$ if $f(\boldsymbol{x}) < +\infty$ and $\partial f(\boldsymbol{x}) \neq \varnothing$.

The following definition formalizes the primal and dual problems relevant to the statement and proof of Theorem 2.1.

Definition S.1.7 (Fenchel Primal and Dual Problems)**.** *Suppose that $f, g : \mathbb{R}^d \to \overline{\mathbb{R}}$ are proper convex functions and consider the primal problem:*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^d} \{f(\boldsymbol{x}) + g(\boldsymbol{x})\}.$$

Then the Fenchel dual problem is:

$$\sup_{\boldsymbol{x}^* \in \mathbb{R}^d} \{-f^*(\boldsymbol{x}^*) - g^*(-\boldsymbol{x}^*)\}.$$

S.1.3 Empirical Process Theory

Assumption 4.1 requires that the class of functions $\mathbf{\Phi}_{\ell b}^{L} := \left\{\varphi_{\ell b}^{L}(\,\cdot\,, \theta, \boldsymbol{\lambda}) : (\theta, \boldsymbol{\lambda}) \in \Theta \times \Lambda\right\}$ satisfy a basic measurability requirement. Pointwise measurability, defined next, is sufficient.

Definition S.1.8. *A class $\mathcal{H}$ of measurable functions on the measurable space $(\mathcal{X}, \mathfrak{X})$ is pointwise measurable if there exists a countable subset $\mathcal{H}' \subset \mathcal{H}$ such that for every $h \in \mathcal{H}$, there exists a sequence $\{h_n\} \in \mathcal{H}'$ with $h_n(\boldsymbol{x}) \to h(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$.*

We now introduce Dudley's entropy condition, which is referenced in Assumptions 4.1 of the main text. Let $||\,\cdot\,||_{Q,2}$ denote the $L_2(Q)$ norm, and let $N(\tau, \mathcal{H}, d)$ denote the smallest number of $d-$balls of radius τ needed to cover $\mathcal{H}$.

Definition S.1.9 (Dudley's Entropy Condition)**.** *Suppose that $\mathcal{H}$ is a collection of real-valued measurable functions on the measurable space $(\mathcal{X}, \mathfrak{X})$ with a measurable envelope function $\boldsymbol{H}$. We say that the class of functions $\mathcal{H}$ satisfies Dudley's entropy condition for the envelope $\boldsymbol{H}$ if:*

$$\int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log N(\varepsilon ||\boldsymbol{H}||_{Q,2}, \mathcal{H}, L_2(Q))}\, d\varepsilon < \infty,$$

where the supremum is taken over all finitely supported probability measures on $(\mathcal{X}, \mathfrak{X})$.

For notational simplicity, let $(\mathbb{W}, \mathscr{W}) = (\mathcal{Y} \times \mathcal{Z}, \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}))$, let $P \in \mathcal{P} \subset \mathscr{P}$, where $\mathscr{P}$ denotes the collection of all probability measures on $\mathscr{W}$. For the results in Section 4, we denote the product probability P^n on the measurable space $(\mathbb{W}^n, \mathscr{W}^n)$ as $\Pr_P$. For the bootstrap results, we take the underlying probability space to be the product of $(\Omega, \mathfrak{F}, \mathbb{P})$ with a probability space $(\Omega^\sharp, \mathfrak{F}^\sharp, \mathbb{P}^\sharp)$ on which we can define the random variables $i_1^{(r)}, \ldots, i_{q_n}^{(r)}$ with uniform distribution on $\{1, \ldots, q_n\}$ for all r. Then take $W_{j,b}^{(r)\sharp} = W_{i_j^{(r)}(\omega^\sharp)}^{(r)}(\omega)$ (see Dudley (2014) p. 324).

S.2 Proofs

S.2.1 Proofs of the Results in the Main Text

Proof of Lemma 2.1. Define:

$$\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) := \left\{ \begin{bmatrix} x_1 \\ \boldsymbol{x}_2 \\ x_3 \\ x_4 \end{bmatrix} \in \mathbb{R}^{d_m+3} : \begin{array}{l} x_1 = \varphi(\boldsymbol{Y}^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ \boldsymbol{x}_2 = \boldsymbol{m}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ x_3 = \delta_1(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ x_4 = \delta_2(\boldsymbol{Y}^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \end{array} \right\}. \quad \text{(S.2.1)}$$

Then:

$$d(\boldsymbol{x}', \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)) = \inf_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)} ||\boldsymbol{x}' - \boldsymbol{x}||$$

$$= \left\| \begin{vmatrix} x_1' - \varphi(\boldsymbol{Y}^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ \boldsymbol{x}_2' - \boldsymbol{m}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ x_3' - \delta_1(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \\ x_4' - \delta_2(\boldsymbol{Y}^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \theta) \end{vmatrix} \right\|,$$

is continuous in $\boldsymbol{x}'$ and measurable in ω under our maintained assumptions. Thus, $(\boldsymbol{x}', \omega) \mapsto d(\boldsymbol{x}', \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta))$ is a Caratheódory function (see Definition S.1.2). Weak measurability of $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)$ follows from Aliprantis and Border (2006) Theorem 18.5. By Lemma S.2.1 and S.2.2, the set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is thus the closure of a countable union of measurable sets, and so weak measurability of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ follows from Aliprantis and Border (2006) Lemma 18.4 and Molchanov (2017) Proposition 1.3.29. Nonemptiness follows from Assumption 2.3(i), which implies the existence of at least one integrable selection (and thus one selection) for each $\theta \in \Theta$ (see Molchanov (2017) p.227). ■

Proof of Proposition 2.1. Fix $\theta \in \Theta^*$, and consider the initial problem:

$$\inf_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathbb{E}[\varphi(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)], \tag{S.2.2}$$

subject to the constraints:

$$\mathbb{E}[\delta_1(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = 0, \qquad \mathbb{E}[\delta_2(\boldsymbol{Y}^\star, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = 0, \qquad \mathbb{E}[\boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \theta)] = \boldsymbol{0}.$$

Recall the set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{Y}^\star, \theta)$ defined in (S.2.1). Then (S.2.2) is equivalent to:

$$\inf_{\substack{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y \\ (X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{Y}^\star, \theta)}} \mathbb{E}[X_1] \text{ s.t. } \mathbb{E}[\boldsymbol{X}_2] = \boldsymbol{0}, \mathbb{E}[X_3] = 0, \ \mathbb{E}[X_4] = 0. \tag{S.2.3}$$

Now define:

$$\mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) := \text{cl}\left(\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) \right).$$

Note that $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{Y}^\star, \theta)$ for some $(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y$ implies that $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}^\circ(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Now suppose $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}^\circ(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. By Lemma S.2.2, there exists a countable collection $\{(\boldsymbol{U}_n, \boldsymbol{Y}_n^\star)\}_{n=1}^\infty$ such that:

$$\text{cl}\left(\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) \right) = \text{cl}\left(\bigcup_{n=1}^\infty \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta) \right),$$

for all $\omega \in \Omega$. Then $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}^\circ(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ implies that for every $k \geq 1$:

$$(X_1(\omega), \boldsymbol{X}_2(\omega), X_3(\omega), X_4(\omega)) \in B_{2^{-k}}(\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta)),$$

for some $n \geq 1$, almost surely. Let $\Omega_{n,k}$ denote the set:

$$\Omega_{n,k} := \left\{ \omega \in \Omega : \left\| \begin{array}{l} X_1'(\omega) - \varphi(\boldsymbol{Y}_n^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \theta) \\ \boldsymbol{X}_2'(\omega) - \boldsymbol{m}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \theta) \\ X_3'(\omega) - \delta_1(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \theta) \\ X_4'(\omega) - \delta_2(\boldsymbol{Y}_n^\star(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \theta) \end{array} \right\| < 2^{-k} \right\}.$$

Under our maintained assumptions this set is measurable. Now set $\Omega_{1,k}' = \Omega_{1,k}$ and $\Omega_{n,k}' = \Omega_{n,k} \setminus \bigcup_{j=1}^{n-1} \Omega_{n,k}'$. Then $\{\Omega_{n,k}'\}_{n=1}^\infty$ is a sequence of disjoint measurable sets with $\mathbb{P}(\bigcup_{n\geq 1} \Omega_{n,k}') = 1$ and:

$$(X_1(\omega), \boldsymbol{X}_2(\omega), X_3(\omega), X_4(\omega)) \in B_{2^{-k}}(\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta)), \quad \forall \omega \in \Omega_{n,k}',$$

for every $k \geq 1$. Let $\Omega_{0,k}' = \Omega \setminus \bigcup_{n\geq 1} \Omega_{n,k}$ and redefine $\Omega_{1,k} \leftarrow \Omega_{0,k}' \cup \Omega_{1,k}'$. Now set:

$$\tilde{\boldsymbol{U}}_k(\omega) = \sum_{n\geq 1} 1_{\Omega_{n,k}}(\omega) \boldsymbol{U}_n(\omega), \qquad \tilde{\boldsymbol{Y}}_k^\star(\omega) = \sum_{n\geq 1} 1_{\Omega_{n,k}}(\omega) \boldsymbol{Y}_n^\star(\omega).$$

Then $(\tilde{\boldsymbol{U}}_k, \tilde{\boldsymbol{Y}}_k^\star) \in \mathcal{M}_u \times \mathcal{M}_y$ and by construction:

$$(X_1(\omega), \boldsymbol{X}_2(\omega), X_3(\omega), X_4(\omega)) \in B(\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \tilde{\boldsymbol{U}}_k(\omega), \tilde{\boldsymbol{Y}}_k^\star(\omega), \theta), 2^{-k}),$$

almost surely. Since this holds for any $k \geq 1$, conclude that $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{Y}^\star, \theta)$ for some $(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y$ if and only if $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}^\circ(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$.

From Lemma S.2.1 we have that $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) = \mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$ for every $\omega \in \Omega$. Combining with the proof above, conclude that $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}^\circ(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ if and only if $(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. Thus (S.2.3) is equivalent to:

$$\inf_{(X_1, \boldsymbol{X}_2, X_3, X_4) \in \text{Sel}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \mathbb{E}[X_1] \text{ subject to } \mathbb{E}[\boldsymbol{X}_2] = 0, \ \mathbb{E}[X_3] = 0 \ \mathbb{E}[X_4] = 0.$$

This completes the proof.

■

Proof of Proposition 2.2. This follows immediately from the definition of the Aumann expectation. ■

Proof of Lemma 2.2. Nonemptiness follows from Molchanov (2017) Theorem 2.1.14, and closedness follows from Molchanov (2017) Theorem 2.1.37. Boundedness follows from integrable boundedness in Assumption 2.3. Convexity follows from Molchanov (2017) Theorem 2.1.26. ■

Proof of Theorem 2.1. Consider the following primal problem:

$$\inf_{\boldsymbol{a}\in A_0\cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)} a_1 = \inf_{\boldsymbol{a}\in\mathbb{R}^{d_m+3}} a_1 + \mathbb{I}\{\boldsymbol{a}\in A_0\cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}, \tag{S.2.4}$$

where $A_0 := \left\{\boldsymbol{x}\in\mathbb{R}^{d_m+3} : x_2 = x_3 = \ldots = x_{d_m+3} = 0\right\}$ and:

$$\mathbb{I}\{\boldsymbol{a}\in A_0\cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\} = \begin{cases} 0, & \text{if } \boldsymbol{a}\in A_0\cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta), \\ +\infty, & \text{otherwise.} \end{cases} \tag{S.2.5}$$

By Lemma 2.2, the set $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ is closed, convex and bounded. The set $A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ is thus closed and convex, being the intersection of two closed and convex sets, and is nonempty since $\theta\in\Theta^*$, by assumption. Claim (i) follows from the Extreme Value Theorem, since (S.2.4) is the minimization of a continuous function over a compact set.

Now the conjugate of the indicator function (S.2.5) is the support function of the set $A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ (Rockafellar (1970) Theorem 13.2); that is:

$$(\mathbb{I}\{\boldsymbol{a}\in\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\})^*(\boldsymbol{a}^*) = \sup_{\boldsymbol{a}\in A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \langle \boldsymbol{a},\boldsymbol{a}^*\rangle = s(\boldsymbol{a}^*, A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)). \tag{S.2.6}$$

Furthermore, direct calculation shows:

$$(a_1)^*(\boldsymbol{a}^*) = \mathbb{I}\{\boldsymbol{a}^* = (1,\mathbf{0},0,0)\} := \begin{cases} 0, & \text{if } \boldsymbol{a}^* = (1,\mathbf{0},0,0), \\ +\infty, & \text{otherwise.} \end{cases} \tag{S.2.7}$$

Using (S.2.6) and (S.2.7), the Fenchel dual problem is (see Definition S.1.7):

$$\sup_{\boldsymbol{a}^*\in\mathbb{R}^{d_m+3}} \{-(a_1)^*(\boldsymbol{a}^*) - s(-\boldsymbol{a}^*, A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta))\} = -s((-1,\mathbf{0},0,0), A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)).$$

From Hörmander (2007) Theorem 2.2.9 we have:

$$\begin{aligned} s(\boldsymbol{a}^*, A_0\cap\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)) &= \text{lsc.}\inf_{\boldsymbol{a}^*=\boldsymbol{a}^{**}+\boldsymbol{a}^{***}} \{s(\boldsymbol{a}^{**}, A_0) + s(\boldsymbol{a}^{***}, \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta))\} \\ &= \text{lsc.}\inf_{\boldsymbol{a}^{**}\in\mathbb{R}^{d_m+3}} \{s(\boldsymbol{a}^{**}, A_0) + s(\boldsymbol{a}^* - \boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta))\}, \end{aligned}$$

where "lsc." denotes the lower semi-continuous regularization.[1] Since A_0 and $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ are nonempty, closed and convex, their support functions are proper, convex, and lower semi-continuous (see Rockafellar (1970) Theorem 13.2, Hörmander (2007) Theorem 2.2.8). Furthermore:

$$\begin{aligned} \text{dom}((s(\boldsymbol{a}^{**}, A_0))^*) &= \text{dom}(\mathbb{I}\{\boldsymbol{a}\in A_0\}) = A_0, \\ \text{dom}((s(\boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)))^*) &= \text{dom}(\mathbb{I}\{\boldsymbol{a}\in\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}) = \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta). \end{aligned}$$

[1] For any convex function $f:\mathbb{R}^d\to\overline{\mathbb{R}}$, the lower semi-continuous regularization of f is the function $g:\mathbb{R}^d\to\overline{\mathbb{R}}$ satisfying $g(x) = \liminf_{y\to x} f(y)$. See Hörmander (2007) p.66.

Note that both $s(\boldsymbol{a}^{**}, A_0))^*$ and $s(\boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)))^*$ are proper, convex and lower-semicontinuous, being the indicator functions of nonempty, convex, and closed sets. Also note that:

$$\operatorname{int}(\operatorname{dom}((s(\boldsymbol{a}^{**}, A_0))^*)) \cap \operatorname{dom}((s(\boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)))^*) = A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta) \neq \varnothing,$$

by assumption. By Lemma S.2.3, we have:

$$\operatorname{ri}(\operatorname{dom}((s(\boldsymbol{a}^{**}, A_0))^*)) \cap \operatorname{ri}(\operatorname{dom}(s(\boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)))^*) \neq \varnothing.$$

From Bauschke, Combettes, et al. (2011) Propositions 6.14 and 15.7—a consequence of the Attouch-Brezis Theorem—we have:

$$\begin{aligned}\operatorname{lsc.} \inf_{\boldsymbol{a}^{**} \in \mathbb{R}^{d_m+3}} &\{s(\boldsymbol{a}^{**}, A_0) + s(\boldsymbol{a}^* - \boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))\} \\ &= \min_{\boldsymbol{a}^{**} \in \mathbb{R}^{d_m+3}} \{s(\boldsymbol{a}^{**}, A_0) + s(\boldsymbol{a}^* - \boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))\}.\end{aligned}$$

That is, the infimal convolution is proper and lower semi-continuous, and the infimum is obtained. Thus:

$$\begin{aligned}
&- s((-1, \mathbf{0}, 0, 0), A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)) \\
&= \max_{\boldsymbol{a}^{**} \in \mathbb{R}^{d_m+3}} \{-s(\boldsymbol{a}^{**}, A_0) - s((-1, \mathbf{0}, 0, 0) - \boldsymbol{a}^{**}, \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))\} \\
&\overset{(1)}{=} \max_{\boldsymbol{a}^{**} \in \mathbb{R}^{d_m+3}} \{-s(\boldsymbol{a}^{**}, A_0) - \mathbb{E}s((-1, \mathbf{0}, 0, 0) - \boldsymbol{a}^{**}, \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta))\} \\
&= \max_{\boldsymbol{a}^{**} \in \mathbb{R}^{d_m+3}} \left\{ - \sup_{\boldsymbol{a} \in A_0} \langle \boldsymbol{a}^{**}, \boldsymbol{a} \rangle - \int \sup_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \{(-1 - a_1^{**})x_1 - \langle \boldsymbol{x}_2, \boldsymbol{a}_2^{**} \rangle - x_3 a_3^{**} - x_4 a_4^{**}\} \, d\mathbb{P} \right\} \\
&\overset{(2)}{=} \max_{(\boldsymbol{a}_2^{**}, a_3^{**}, a_4^{**}) \in \mathbb{R}^{d_m+2}} \left\{ - \int \sup_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \{-x_1 - \langle \boldsymbol{x}_2, \boldsymbol{a}_2^{**} \rangle - x_3 a_3^{**} - x_4 a_4^{**}\} \, d\mathbb{P} \right\} \\
&= \max_{(\boldsymbol{a}_2^{**}, a_3^{**}, a_4^{**}) \in \mathbb{R}^{d_m+2}} \int \inf_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \{x_1 + \langle \boldsymbol{x}_2, \boldsymbol{a}_2^{**} \rangle + x_3 a_3^{**} + x_4 a_4^{**}\} \, d\mathbb{P} \\
&\overset{(3)}{=} \max_{\boldsymbol{a}_2^{**} \in \mathbb{R}^{d_m}} \int \inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \inf_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} \{\varphi(\boldsymbol{y}^\star, \boldsymbol{Z}, \boldsymbol{u}, \theta) + \langle \boldsymbol{a}_2^{**}, \boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \theta) \rangle\} \, d\mathbb{P}.
\end{aligned}$$

Equality (1) follows from Molchanov (2017) Theorem 2.1.35 after noting that $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is an integrably bounded random closed set, that the probability space is non-atomic by Assumption 2.1, and that $\mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is closed (and thus equal to the selection expectation) by Molchanov (2017) Theorem 2.1.37. Equality (2) follows after noting:

$$\sup_{\boldsymbol{a} \in A_0} \langle \boldsymbol{a}^{**}, \boldsymbol{a} \rangle = \begin{cases} 0, & \text{if } a_1^{**} = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Finally, equality (3) follows from the definition of $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$. This proves claim (ii) in the state-

ment of Theorem 2.1. For claim (iii), note that:

$$\operatorname{dom}(a_1) = \mathbb{R}^{d_m+3}, \qquad \operatorname{dom}(\mathbb{I}\{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}) = A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta),$$

where the second equality follows from boundedness of $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ by Assumption 2.3. Note that $\boldsymbol{a} \mapsto a_1$ and $\boldsymbol{a} \mapsto \mathbb{I}\{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}$ are proper and convex. Also note that:

$$\operatorname{int}(\operatorname{dom}(a_1)) \cap \operatorname{dom}(\mathbb{I}\{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}) = A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta) \neq \varnothing,$$

since $\theta \in \Theta^*$. By Lemma S.2.3, we have:

$$\operatorname{ri}(\operatorname{dom}(a_1)) \cap \operatorname{ri}(\operatorname{dom}(\mathbb{I}\{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\})) \neq \varnothing. \tag{S.2.8}$$

Since $\boldsymbol{a} \mapsto a_1$ and $\boldsymbol{a} \mapsto \mathbb{I}\{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)\}$ are proper and convex, claim (iii) follows from (S.2.8) and Theorem 31.1(a) in Rockafellar (1970).

■

Proof of Theorem 2.2. By Theorem 2.1, we have $\overline{\operatorname{conv}}(\Phi^*) = [\varphi'_{\ell b}, \varphi'_{ub}]$, where:

$$\varphi'_{\ell b} = \inf_{\theta\in\Theta^*} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \int \inf_{u\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \inf_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} \{\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta)\rangle\}\, d\mathbb{P},$$
$$\varphi'_{ub} = \sup_{\theta\in\Theta^*} \inf_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \int \sup_{u\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} \{\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta)\rangle\}\, d\mathbb{P}.$$

The claim follows if $\varphi'_{\ell b}$ and φ'_{ub} can be replaced with $\varphi_{\ell b}$ and φ_{ub}. Focusing on $\varphi_{\ell b}$, for any $S \subset \Theta$ define:

$$\varphi_{\ell b}[S] = \inf_{\theta\in S} \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \int \inf_{u\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \inf_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} \{\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta)\rangle\}\, d\mathbb{P}.$$

Now suppose by way of contradiction that $\varphi_{\ell b}[\Theta] < \varphi_{\ell b}[\Theta^*]$. Since:

$$\varphi_{\ell b}[\Theta] = \min\left\{\varphi_{\ell b}[\Theta \setminus \Theta^*], \varphi_{\ell b}[\Theta^*]\right\},$$

conclude that $\varphi_{\ell b}[\Theta \setminus \Theta^*] < \varphi_{\ell b}[\Theta^*]$. By definition of the infimum, for every $\varepsilon > 0$ there exists a $\theta^\dagger \in \Theta \setminus \Theta^*$ such that:

$$\varphi_{\ell b}[\theta^\dagger] \leq \varphi_{\ell b}[\Theta \setminus \Theta^*] + \varepsilon.$$

Now set $\varepsilon = \varphi_{\ell b}[\Theta^*] - \varphi_{\ell b}[\Theta \setminus \Theta^*]$ and choose the corresponding $\theta^\dagger \in \Theta \setminus \Theta^*$ so that:

$$\varphi_{\ell b}[\theta^\dagger] \leq \varphi_{\ell b}[\Theta \setminus \Theta^*] + \varepsilon = \varphi_{\ell b}[\Theta^*].$$

But since $\theta^\dagger \notin \Theta^*$, Lemma S.2.4 implies that $\varphi_{\ell b}[\theta^\dagger] = +\infty$. Thus, the previous display can hold only if $\varphi_{\ell b}[\Theta^*] = +\infty$. But for any fixed $\check{\theta} \in \Theta^*$, Lemma S.2.4 implies $\varphi_{\ell b}[\Theta^*] \leq \varphi_{\ell b}[\check{\theta}] < \infty$, a

contradiction. Thus, conclude that $\varphi'_{\ell b} = \varphi_{\ell b}$. Since a similar proof holds for φ'_{ub}, this completes the proof.

■

Proof of Proposition 2.3. If Assumption 2.4 holds, then there exists a random vector $\boldsymbol{U}' : \Omega \to \mathcal{U} \subset \mathbb{R}^{d_u}$ such that $(\mathbb{E}[||\boldsymbol{U} - \boldsymbol{U}'||_q^q])^{1/q} \leq \rho$. Taking π^* as the joint distribution of $\boldsymbol{U}$ and $\boldsymbol{U}'$, we have:

$$W_q(P_{\boldsymbol{U}}, P_{\boldsymbol{U}'}) \leq \left(\int ||\boldsymbol{u} - \boldsymbol{u}'||_q^q \, d\pi^* \right)^{1/q} \leq \rho.$$

This proves (i). Part (ii) follows directly from the definition of the $q-$Wasserstein distance as the minimizer over all couplings $(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{U}}')$ of $(\boldsymbol{U}, \boldsymbol{U}')$, and from the fact that the infimum in $W_q(P_{\boldsymbol{U}}, P_{\boldsymbol{U}'})$ is attained (see Villani (2003) Theorem 1.3). For part (iii), Champion and De Pascale (2011) Theorem 1.1 (and 5.1) imply the existence of a Borel measurable function $T : \mathcal{U} \to \mathcal{U}$ (i.e. a transport map), defined up to null sets of $P_{\boldsymbol{U}}$, such that $(\boldsymbol{U}, T(\boldsymbol{U}))$ is a (possibly non-unique) solution to the problem (2.15). The claim follows by setting $\boldsymbol{U}' = T(\boldsymbol{U})$. ■

For Theorem 2.3, we require some modifications to the assumptions in Section 2.1 and Section 2.2. Note that the inequality in Assumption 2.4 can be equivalently written as:

$$\mathbb{E}[m_{0,\rho}(\boldsymbol{U}, \boldsymbol{U}', \kappa)] = 0, \text{ where } m_{0,\rho}(\boldsymbol{u}, \boldsymbol{u}', \kappa) := ||\boldsymbol{u} - \boldsymbol{u}'||_q^q - \rho^q + \kappa,$$

for some slackness variable $\kappa \geq 0$. In order words, Assumption 2.4 can be imposed in our framework by adding an additional moment condition. Now consider the following revised random set:

$$\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa) := \text{cl} \left\{ (x_1, x_2, \boldsymbol{x}_3, x_4, x_5) \in \mathbb{R}^{d_m+4} : \begin{array}{ll} x_1 = \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta), & \\ x_2 = m_{0,\rho}(\boldsymbol{u}, \boldsymbol{U}', \kappa), & \\ \boldsymbol{x}_3 = \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta), & (\boldsymbol{u}, \boldsymbol{y}^\star) \in \mathcal{U} \times \mathcal{Y}, \\ x_4 = \delta_1(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta), & \\ x_5 = \delta_2(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta), & \end{array} \right\}.$$

Furthermore, we require the following revised version of Assumption 2.3, replacing the random set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ from (2.3) with the new random set $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)$.

Assumption S.2.1. *(i) There exists a* $\kappa \in \mathbb{R}_+$ *such that* $\mathbb{E}[m_{0,\rho}(\boldsymbol{U}, \boldsymbol{U}', \kappa)] = 0$. *(ii) For every* $(\theta, \kappa) \in \Theta \times \mathbb{R}_+$, $\mathbb{E}\left[\sup\{||\boldsymbol{X}|| : \boldsymbol{X} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)\}\right] < \infty$.

Definition 2.1 can be revised to accommodate Assumption 2.4 and S.2.1 by adding the condition that $\mathbb{E}[m_{0,\rho}(\widetilde{\boldsymbol{U}}, \boldsymbol{U}', \kappa)] = 0$ must be satisfied for some $\kappa \geq 0$. Also note that Lemma 2.1, Proposition

2.1, Proposition 2.2 and Lemma 2.2 can all be revised accordingly, and their proofs remain mostly unchanged.

Proof of Theorem 2.3. The proof is nearly identical to the proofs of Theorem 2.1 and Theorem 2.2, with the exception that the form of the dual problem is slightly different. Here we highlight the few key differences. Let E^* denote the set of all pairs (θ, κ) satisfying all conditions of the revised version of Definition 2.1 (i.e. the "joint identified set" for (θ, κ)). Then following the proof of Theorem 2.1, we obtain for any $(\theta, \kappa) \in E^*$:

$$\inf_{\boldsymbol{a} \in A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)} a_1 = -s((-1, 0, \mathbf{0}, 0, 0), A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)).$$

Simplifying, we have:

$$\begin{aligned}
&- s((-1, 0, \mathbf{0}, 0, 0), A_0 \cap \mathbf{E}\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)) \\
&= \max_{(a_2^{**}, \boldsymbol{a}_3^{**}, a_4^{**}, a_5^{**}) \in \mathbb{R}_+ \times \mathbb{R}^{d_m + 2}} \int \inf_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{U}', \theta, \kappa)} \{x_1 + a_2^{**} x_2 + \langle \boldsymbol{x}_3, \boldsymbol{a}_3^{**} \rangle + x_4 a_4^{**} + x_5 a_5^{**}\} \, d\mathbb{P} \\
&= \max_{(a_2^{**}, \boldsymbol{a}_3^{**}) \in \mathbb{R}_+ \times \mathbb{R}^{d_m}} a_2^{**} \kappa + \int \inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \inf_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} \{\varphi(\boldsymbol{y}^\star, \boldsymbol{Z}, \boldsymbol{u}, \theta) \\
&\qquad + a_2^{**}(||\boldsymbol{u} - \boldsymbol{U}'||_q^q - \rho^q) + \langle \boldsymbol{a}_3^{**}, \boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \theta) \rangle\} \, d\mathbb{P}.
\end{aligned}$$

Thus the lower bound is:

$$\begin{aligned}
&\inf_{(\theta, \kappa) \in E^*} \max_{(a_2^{**}, \boldsymbol{a}_3^{**}) \in \mathbb{R}_+ \times \mathbb{R}^{d_m}} a_2^{**} \kappa + \int \inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \inf_{\boldsymbol{y}^\star \in \mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{u}, \theta)} \{\varphi(\boldsymbol{y}^\star, \boldsymbol{Z}, \boldsymbol{u}, \theta) \\
&\qquad + a_2^{**}(||\boldsymbol{u} - \boldsymbol{U}'||_q^q - \rho^q) + \langle \boldsymbol{a}_3^{**}, \boldsymbol{m}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{u}, \theta) \rangle\} \, d\mathbb{P}.
\end{aligned}$$

A similar proof to the proof of Theorem 2.2 shows that E^* can be replaced with $\Theta \times \mathbb{R}_+$. Since $\kappa \in \mathbb{R}_+$ and $a_2^{**} \in \mathbb{R}_+$, the infimum over $\kappa \in \mathbb{R}_+$ is always obtained at $\kappa = 0$. ■

Proof of Lemma 3.1. Throughout the proof we fix $\theta \in \Theta$. Since $\boldsymbol{\lambda} \mapsto -\hat{\varphi}_{lb}^L(\theta, \boldsymbol{\lambda})$ is proper (since $\mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ and $\mathcal{Y}^\star(\boldsymbol{Z}, \boldsymbol{U}, \theta)$ are nonempty) and convex, nonemptiness of the subdifferential of $\boldsymbol{\lambda} \mapsto -\hat{\varphi}_{lb}^L(\theta, \boldsymbol{\lambda})$ at every $\boldsymbol{\lambda} \in \text{ri}(\text{dom}(-\hat{\varphi}_{lb}^L(\theta, \boldsymbol{\lambda})))$ follows from Rockafellar (1970) Theorem 23.4. The fact that the subdifferential is convex and closed follows immediately from Definition S.1.6, since it is the intersection of closed halfspaces (see Rockafellar (1970) p.215). For the second claim, we iteratively apply rules for computing subdifferentials. In particular, consider the following rules:

(i) For any $\alpha \geq 0$, we have $\partial(\alpha f)(\bar{\boldsymbol{x}}) = \alpha \partial f(\bar{\boldsymbol{x}}) := \{\alpha \boldsymbol{g} : \boldsymbol{g} \in \partial f(\bar{\boldsymbol{x}})\}$. This follows immediately from the definition of the subgradient.

(ii) If $f_i(\boldsymbol{x})$ is a proper, convex function for $i = 1, \ldots n$, and $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + \ldots + f_n(\boldsymbol{x})$, then $\partial f_1(\boldsymbol{x}) + \ldots + \partial f_n(\boldsymbol{x}) \subset \partial f(\boldsymbol{x}) \; \forall \boldsymbol{x} \in \mathbb{R}^n$. See Theorem 23.8 in Rockafellar (1970).

(iii) Suppose that $f(\boldsymbol{x}) = \sup_{\boldsymbol{z}\in Z} \phi(\boldsymbol{x}, \boldsymbol{z})$ where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{z} \in \mathbb{R}^m$, $\phi : \mathbb{R}^n \times \mathbb{R}^m \to (-\infty, \infty]$, where $Z \subset \mathbb{R}^m$ is compact. Suppose that $\phi(\,\cdot\,, \boldsymbol{z})$ is proper, convex, and lower semi-continuous for each $\boldsymbol{z}$. Also suppose that $\text{int}(\text{dom}(f)) \neq \varnothing$, and that ϕ is continuous on $\text{int}(\text{dom}(f)) \times \mathbb{R}^m$. Then for any $\boldsymbol{x} \in \text{int}(\text{dom}(f))$, if $\boldsymbol{z}^*$ obtains the supremum, then $\partial_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{z}^*) \subset \partial_{\boldsymbol{x}} f(\boldsymbol{x})$. This is an extension of Danskin's Theorem first proved by Bertsekas (1971) (Proposition A.22).

From the definition of $\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda})$ we have:

$$-\hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} \sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \Big(-\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle \Big). \tag{S.2.9}$$

Now note that for each $(\boldsymbol{u}, \boldsymbol{y}^\star)$ the function:

$$\boldsymbol{\lambda} \mapsto -\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle, \tag{S.2.10}$$

is proper, convex, lower semi-continuous and differentiable with (sub)gradient given by $\boldsymbol{g}_i(\boldsymbol{u}) := -\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}_i, \theta)$ for every $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^{d_m}$. By continuity of the objective function and compactness of $\mathcal{S}_i$, the supremum in (S.2.9) is obtained. Set $\boldsymbol{g}_i^* := \boldsymbol{g}_i(\boldsymbol{u}_i^*)$ and apply rule (iii) above to conclude that:

$$\boldsymbol{g}_i^* \in \partial_{\boldsymbol{\lambda}} \left(\sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \Big(-\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \bar{\boldsymbol{\lambda}}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle \Big) \right). \tag{S.2.11}$$

Now applying rules (i) and (ii) above, we have:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \partial_{\boldsymbol{\lambda}} \left(\sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \Big(-\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \bar{\boldsymbol{\lambda}}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle \Big) \right) \\ &\quad \subset \partial_{\boldsymbol{\lambda}} \left(\frac{1}{n}\sum_{i=1}^{n} \sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \Big(-\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \bar{\boldsymbol{\lambda}}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle \Big) \right). \end{aligned} \tag{S.2.12}$$

Finally, combining (S.2.11) and (S.2.12), conclude that:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^* \in \partial_{\boldsymbol{\lambda}} \left(\frac{1}{n}\sum_{i=1}^{n} \sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{y}_i,\boldsymbol{z}_i,\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{z}_i,\boldsymbol{u},\theta)} \Big(-\varphi(\boldsymbol{y}^\star, \boldsymbol{z}_i, \boldsymbol{u}, \theta) - \langle \bar{\boldsymbol{\lambda}}, \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{u}, \theta)\rangle \Big) \right).$$

This completes the proof of the second claim. The final claim follows from Danskin's Theorem (e.g. Bertsekas, Nedic, and Ozdaglar (2003) Proposition 4.5.1) after noting that (S.2.10) is differentiable in $\boldsymbol{\lambda}$ for each $(\boldsymbol{u}, \boldsymbol{y}^\star)$. ■

Proof of Theorem 4.1. Suppose without loss of generality that:

$$\inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] \geq \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}).$$

Now fix any $\varepsilon > 0$ and let θ' be any value satisfying:

$$\sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta', \boldsymbol{\lambda}) \leq \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) + \varepsilon.$$

Then:

$$\begin{aligned}
&\left| \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \right| \\
&\quad = \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \\
&\quad \leq \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta', \boldsymbol{\lambda})] - \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta', \boldsymbol{\lambda}) + \varepsilon \\
&\quad \leq \sup_{\theta\in\Theta} \left(\sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \right) + \varepsilon \\
&\quad \leq \sup_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \left| \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \right| + \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, conclude that:

$$\left| \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \inf_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \right| \leq \sup_{\theta\in\Theta} \sup_{\boldsymbol{\lambda}\in\Lambda} \left| \mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \theta, \boldsymbol{\lambda})] - \hat{\varphi}^L_{\ell b}(\theta, \boldsymbol{\lambda}) \right|.$$

The result then follows from Lemma S.2.6. ■

Proof of Theorem 4.2. Assumption 4.1 matches Assumption 3.1 in Marcoux, Russell, and Wan (2023), with the exception of the uniform over $P \in \mathcal{P}$ moment condition in Assumption 3.1(ii) needed for their uniformity results. Furthermore Assumption 4.2 matches Assumption 3.3 in Marcoux, Russell, and Wan (2023), with the exception of the uniform over $P \in \mathcal{P}$ condition in Assumption 3.3(i) which is also needed for their uniformity results. Assumptions 3.2 and 3.4 of Marcoux, Russell, and Wan (2023) are also trivially satisfied in our context since our test statistic involves no conditional moments. The first claim then follows from Theorem 3.1 in Marcoux, Russell, and Wan (2023), and the second claim follows from Theorem 3.2 in Marcoux, Russell, and Wan (2023). ■

S.2.2 Proofs of Supporting Results

Lemma S.2.1. *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Then:*

$$\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) = cl\left(\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star)\in\mathcal{M}_u\times\mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) \right), \tag{S.2.13}$$

for every $\omega \in \Omega$, where $\mathcal{X}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ is the random set from (2.3) and $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)$ is the random set from (S.2.1).

Proof of Lemma S.2.1. Fix any $\omega \in \Omega$ and let $\mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$ denote the set on the right side of (S.2.13). By definition of $\mathcal{M}_u$ and $\mathcal{M}_y$, $\boldsymbol{U} \in \mathcal{M}_u$ has range $\mathcal{U}$ and $\boldsymbol{Y}^\star \in \mathcal{M}_y$ has range $\mathcal{Y}$.

Thus, $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) \subset \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$ for every $(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y$, and so we must have $\mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) \subset \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$. Now noting that the maps $\boldsymbol{U}(\omega) = \boldsymbol{u}$ and $\boldsymbol{Y}^\star(\omega) = \boldsymbol{y}^\star$ are measurable for any $\boldsymbol{u} \in \mathcal{U}$ and $\boldsymbol{y}^\star \in \mathcal{Y}$, we have:

$$\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) = \bigcup_{(\boldsymbol{u}, \boldsymbol{y}^\star) \in \mathcal{U} \times \mathcal{Y}} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{u}, \boldsymbol{y}^\star, \theta) \subset \mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta).$$

Since the same proof applies for any $\omega \in \Omega$, conclude that $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta) = \mathcal{X}^\circ(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \theta)$ for every $\omega \in \Omega$. ■

Lemma S.2.2. *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Then there exists a countable collection* $\{(\boldsymbol{U}_n, \boldsymbol{Y}_n^\star)\}_{n=1}^\infty$ *of random vectors such that:*

$$cl\left(\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)\right) = cl\left(\bigcup_{n=1}^\infty \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta)\right),$$

for all $\omega \in \Omega$, *where* $\mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta)$ *is the random set from* (S.2.1).

Proof of Lemma S.2.2. Since random vectors are measurable by definition, for any countable collection $\{(\boldsymbol{U}_n, \boldsymbol{Y}_n^\star)\}_{n=1}^\infty$ we have:

$$\bigcup_{n=1}^\infty \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta) \subset \bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta).$$

Now suppose by way of contradiction that for every countable collection $\{(\boldsymbol{U}_n, \boldsymbol{Y}_n^\star)\}_{n=1}^\infty$ of random vectors we have:

$$\left(\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \boldsymbol{Y}^\star(\tilde{\omega}), \theta)\right) \setminus \mathrm{cl}\left(\bigcup_{n=1}^\infty \mathcal{X}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_n(\tilde{\omega}), \boldsymbol{Y}_n^\star(\tilde{\omega}), \theta)\right) \neq \varnothing, \tag{S.2.14}$$

for some $\tilde{\omega} \in \Omega$.

By Assumption 2.1, the set $\mathcal{Y} \times \mathcal{Z}$ is countable, so let $\{(\boldsymbol{y}_k, \boldsymbol{z}_k)\}_{k=1}^\infty$ denote an enumeration of this set. For each $k \geq 1$ there exists a partition of $\mathcal{U} \times \mathcal{Y}$ into four sets $\{A_{k,j}\}_{j=1}^4$ such that $(\boldsymbol{u}, \boldsymbol{y}^\star) \mapsto (\delta_1(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta), \delta_2(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta))$ is constant for all $(\boldsymbol{u}, \boldsymbol{y}^\star) \in A_{k,j}$. Now let $\{(\boldsymbol{u}_{k,j,\ell}, \boldsymbol{y}_{k,j,\ell}^\star)\}_{\ell=1}^\infty \subset A_{k,j}$ be a dense subset of $A_{k,j}$ for each $k \geq 1$ and $1 \leq j \leq 4$, and let $\{(\boldsymbol{u}_{k,j,\ell}, \boldsymbol{y}_{k,j,\ell}^\star) : \ell, k \geq 1, 1 \leq j \leq 4\}$ be the enumeration of all points constructed in this way. After relabeling, the collection $\{(\boldsymbol{u}_{k,j,\ell}, \boldsymbol{y}_{k,j,\ell}^\star) : \ell, k \geq 1, 1 \leq j \leq 4\}$ can be rewritten as $\{(\boldsymbol{u}_n, \boldsymbol{y}_n^\star) : n \geq 1\}$.

Let $\{(\boldsymbol{U}_n, \boldsymbol{Y}_n^\star)\}_{n=1}^\infty$ be a collection of constant random variables such that $(\boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega)) = (\boldsymbol{u}_n, \boldsymbol{y}_n^\star)$ for all $\omega \in \Omega$, for each $n \geq 1$. For this countable collection, let $(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y$ be the corresponding pair satisfying (S.2.14), and let $\boldsymbol{x}^*(\tilde{\omega})$ be any point lying in the set in (S.2.14).

Then from (S.2.14), there exists an open set V containing $\boldsymbol{x}^*(\tilde{\omega})$ such that:

$$V \cap \text{cl}\left(\bigcup_{n=1}^{\infty} \mathcal{X}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_n(\tilde{\omega}), \boldsymbol{Y}_n^\star(\tilde{\omega}), \theta)\right) = \varnothing. \tag{S.2.15}$$

By definition of $\mathcal{X}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \boldsymbol{Y}^\star(\tilde{\omega}), \theta)$, the point $\boldsymbol{x}^*(\tilde{\omega})$ is of the form:

$$\boldsymbol{x}^*(\tilde{\omega}) = \begin{bmatrix} \boldsymbol{x}_1^*(\tilde{\omega}) \\ \boldsymbol{x}_2^*(\tilde{\omega}) \end{bmatrix}, \quad \boldsymbol{x}_1^*(\tilde{\omega}) = \begin{bmatrix} \varphi(\boldsymbol{Y}^\star(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \theta) \\ \boldsymbol{m}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \theta) \end{bmatrix}, \quad \boldsymbol{x}_2^*(\tilde{\omega}) = \begin{bmatrix} \delta_1(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \theta) \\ \delta_2(\boldsymbol{Y}^\star(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}(\tilde{\omega}), \theta) \end{bmatrix}.$$

By Assumption 2.3, $(\boldsymbol{U}(\tilde{\omega}), \boldsymbol{Y}^\star(\tilde{\omega})) \in A_{k,j}$ for some $k, j \geq 1$, and:

$$\boldsymbol{x}_2^*(\tilde{\omega}) = \begin{bmatrix} \delta_1(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{u}, \theta) \\ \delta_2(\boldsymbol{y}^\star, \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{u}, \theta) \end{bmatrix}, \qquad \forall (\boldsymbol{u}, \boldsymbol{y}^\star) \in A_{k,j}.$$

By construction, there exists a subset of $\{(\boldsymbol{U}_n(\tilde{\omega}), \boldsymbol{Y}_n^\star(\tilde{\omega}))\}_{n=1}^{\infty}$ that is dense in $A_{k,j}$, and thus, there exists a subsequence $\{(\boldsymbol{U}_{n_m}(\tilde{\omega}), \boldsymbol{Y}_{n_m}^\star(\tilde{\omega}))\}_{m=1}^{\infty} \subset A_n$ such that $(\boldsymbol{U}_{n_m}(\tilde{\omega}), \boldsymbol{Y}_{n_m}^\star(\tilde{\omega})) \to (\boldsymbol{U}(\tilde{\omega}), \boldsymbol{Y}^\star(\tilde{\omega}))$. By Assumptions 2.1 and 2.2, the maps $(\boldsymbol{y}^\star, \boldsymbol{u}) \mapsto \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta)$ and $\boldsymbol{u} \mapsto \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta)$ are continuous. Thus, along the subsequence $\{(\boldsymbol{U}_{n_m}(\tilde{\omega}), \boldsymbol{Y}_{n_m}^\star(\tilde{\omega}))\}_{m=1}^{\infty}$ we have:

$$\begin{bmatrix} \varphi(\boldsymbol{Y}_{n_m}^\star(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_{n_m}(\tilde{\omega}), \theta) \\ \boldsymbol{m}(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_{n_m}(\tilde{\omega}), \theta) \\ \delta_1(\boldsymbol{Y}(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_{n_m}(\tilde{\omega}), \theta) \\ \delta_2(\boldsymbol{Y}_{n_m}^\star(\tilde{\omega}), \boldsymbol{Z}(\tilde{\omega}), \boldsymbol{U}_{n_m}(\tilde{\omega}), \theta) \end{bmatrix} \to \boldsymbol{x}^*(\tilde{\omega}).$$

But this contradicts (S.2.15). Conclude that:

$$\bigcup_{(\boldsymbol{U}, \boldsymbol{Y}^\star) \in \mathcal{M}_u \times \mathcal{M}_y} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}(\omega), \boldsymbol{Y}^\star(\omega), \theta) \subset \text{cl}\left(\bigcup_{n=1}^{\infty} \mathcal{X}(\boldsymbol{Y}(\omega), \boldsymbol{Z}(\omega), \boldsymbol{U}_n(\omega), \boldsymbol{Y}_n^\star(\omega), \theta)\right),$$

for every $\omega \in \Omega$, which completes the proof. ■

Lemma S.2.3. *Suppose that $f, g : \mathbb{R}^d \to (-\infty, +\infty]$ are proper convex functions, and suppose that either $\text{int}(\text{dom}(f)) \cap \text{dom}(g) \neq \varnothing$ or $\text{dom}(f) \cap \text{int}(\text{dom}(g)) \neq \varnothing$. Then $\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \varnothing$.*

Proof of Lemma S.2.3. The proof is essentially from the proof of Bauschke, Combettes, et al. (2011) Proposition 6.19. In particular, suppose that $\boldsymbol{x} \in \text{int}(\text{dom}(f)) \cap \text{dom}(g)$. Then $B(\boldsymbol{x}, \delta) := \{\boldsymbol{y} \in \mathbb{R}^d : ||\boldsymbol{y} - \boldsymbol{x}|| \leq \delta\} \subset \text{int}(\text{dom}(f))$ for some $\delta > 0$. But then $B(\boldsymbol{0}, \delta) = \boldsymbol{x} - B(\boldsymbol{x}, \delta) \subset \text{dom}(g) - \text{int}(\text{dom}(f))$ and therefore $\boldsymbol{0} \in \text{int}(\text{dom}(g) - \text{dom}(f))$. This implies $\boldsymbol{0} \in \text{ri}(\text{dom}(g) - \text{dom}(f))$. By Rockafellar (1970) Corollary 6.6.2, $\text{ri}(\text{dom}(f) - \text{dom}(g)) = \text{ri}(\text{dom}(f)) - \text{ri}(\text{dom}(g))$, so that $\boldsymbol{0} \in \text{ri}(\text{dom}(f)) - \text{ri}(\text{dom}(g))$. But this is true if and only if $\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \varnothing$. The same proof follows under the condition $\text{dom}(f) \cap \text{int}(\text{dom}(g)) \neq \varnothing$ by interchanging the roles of f and g. ■

Lemma S.2.4. *Suppose Assumptions 2.1, 2.2 and 2.3 hold, and define:*

$$\varphi_{\ell b}(\theta) := \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \int \inf_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \inf_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} \{\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta) + \langle\boldsymbol{\lambda},\boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta)\rangle\}\, d\mathbb{P},$$

$$\varphi_{ub}(\theta) := \inf_{\boldsymbol{\lambda}\in\mathbb{R}^{d_m}} \int \sup_{\boldsymbol{u}\in\mathcal{U}(\boldsymbol{Y},\boldsymbol{Z},\theta)} \sup_{\boldsymbol{y}^\star\in\mathcal{Y}^\star(\boldsymbol{Z},\boldsymbol{u},\theta)} \{\varphi(\boldsymbol{y}^\star,\boldsymbol{Z},\boldsymbol{u},\theta) + \langle\boldsymbol{\lambda},\boldsymbol{m}(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{u},\theta)\rangle\}\, d\mathbb{P}.$$

Then $\varphi_{\ell b}(\theta) < \infty$ for every $\theta \in \Theta^*$, and $\varphi_{\ell b}(\theta) = \infty$ for every $\theta \notin \Theta^*$. Furthermore, $\varphi_{ub}(\theta) > -\infty$ for every $\theta \in \Theta^*$, and $\varphi_{ub}(\theta) = -\infty$ for every $\theta \notin \Theta^*$.

Proof of Lemma S.2.4. Focus on $\varphi_{\ell b}(\theta)$. By Lemma 2.1, the set $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ is compact. Inspecting the primal problem (2.10), the result follows from the fact that $\mathbf{E}\mathcal{X}(\boldsymbol{Y},\boldsymbol{Z},\theta)$ has nonempty intersection with the hyperplane $A_0 = \{\boldsymbol{x} \in \mathbb{R}^{d_m+3} : x_2 = \ldots = x_{d_m+3} = 0\}$ if and only if $\theta \in \Theta^*$, which follows in turn from Definition 2.1. ■

Lemma S.2.5. *Suppose that Assumptions 4.1 and 4.2 hold, and suppose that $\epsilon_n = o(n^{-1/2})$, where $\{\epsilon_n\}_{n\geq 1}$ is from (4.4). Then for any $P \in \mathcal{P}_\tau$:*

$$\lim_{M\to\infty} \limsup_{n\to\infty} Pr_P\left(\sup_{\boldsymbol{\lambda}\in\Lambda} \frac{\sqrt{n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\hat{\theta}_n,\boldsymbol{\lambda})] - \tau)}{\varsigma_P(\hat{\theta}_n)} > M\right) = 0.$$

Proof of Lemma S.2.5. As discussed in the proof of Theorem 4.2, Assumptions 4.1 and 4.2 match the Assumptions 3.1 and 3.3 in Marcoux, Russell, and Wan (2023), and their Assumption 3.3 is trivially satisfied in our setting. Thus the result follows from Marcoux, Russell, and Wan (2023) Lemma 3.1. ■

Lemma S.2.6. *Suppose that Assumptions 4.1 and 4.2 hold. Then:*

$$\lim_{M\to\infty} \limsup_{n\to\infty} Pr_P\left(\sup_{\theta\in\Theta}\sup_{\boldsymbol{\lambda}\in\Lambda} \left|\sqrt{n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\theta,\boldsymbol{\lambda})] - \hat{\varphi}^L_{\ell b}(\theta,\boldsymbol{\lambda}))\right| > M\right) = 0.$$

Proof of Lemma S.2.6. Define:

$$||\mathbb{G}_{n,P}||_{\boldsymbol{\Phi}^L_{\ell b}} := \sup_{\theta\in\Theta}\sup_{\boldsymbol{\lambda}\in\Lambda} \left|\sqrt{n}(\mathbb{E}_P[\varphi^L_{\ell b}(\boldsymbol{Y}_i,\boldsymbol{Z}_i,\theta,\boldsymbol{\lambda})] - \hat{\varphi}^L_{\ell b}(\theta,\boldsymbol{\lambda}))\right|.$$

It suffices to show that for any $\varepsilon > 0$ there exists an $M > 0$ and an N such that:

$$\sup_{n\geq N} \Pr_P\left(||\mathbb{G}_{n,P}||_{\boldsymbol{\Phi}^L_{\ell b}} > M\right) < \varepsilon. \tag{S.2.16}$$

By Markov's inequality, we have:

$$\Pr_P\left(||\mathbb{G}_{n,P}||_{\boldsymbol{\Phi}^L_{\ell b}} > M\right) \leq \frac{\mathbb{E}_P\,||\mathbb{G}_{n,P}||_{\boldsymbol{\Phi}^L_{\ell b}}}{M}. \tag{S.2.17}$$

The remainder of the proof will show that $\mathbb{E}_P\,||\mathbb{G}_{n,P}||_{\boldsymbol{\Phi}^L_{\ell b}}$ is bounded above by a constant that

is independent of n. To this end, note that under Assumption 4.1, the class $\mathbf{\Phi}_{\ell b}^L$ is pointwise measurable and satisfies:

$$\int_0^\infty \sup_{Q\in\mathcal{Q}} \sqrt{\log N(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, L_2(Q))}\, d\epsilon$$
$$= \int_0^1 \sup_{Q\in\mathcal{Q}} \sqrt{\log N(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, L_2(Q))}\, d\epsilon < \infty, \tag{S.2.18}$$

with the supremum taken over all probability measures with finite support. Now define:

$$J(\delta, \mathbf{\Phi}_{\ell b}^L) := \sup_{Q\in\mathcal{Q}} \int_0^\delta \sqrt{1 + \log N(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, ||\cdot||_{Q,2})}\, d\epsilon.$$

Then (S.2.18) implies that $J(1, \mathbf{\Phi}_{\ell b}^L) < \infty$. By Theorem 2.14.1 in Van Der Vaart and Wellner (1996), we have:

$$\mathbb{E}_P ||\mathbb{G}_{n,P}||_{\mathbf{\Phi}_{\ell b}^L} \leq C' J(1, \mathbf{\Phi}_{\ell b}^L) ||\overline{M}||_{P,2}, \tag{S.2.19}$$

for some finite constant C'. By Assumption 4.1, there exists an $\delta > 0$ such that $\mathbb{E}_P[\overline{M}(\boldsymbol{Y}_i, \boldsymbol{Z}_i)^{2+\delta}] \leq C$ for some $C < \infty$, so that by Holder's inequality $||\overline{M}||_{P,2} \leq (\mathbb{E}_P[\overline{M}(\boldsymbol{Y}_i, \boldsymbol{Z}_i)^{2+\delta}])^{2/(2+\delta)} \leq C^{2/(2+\delta)}$. Thus, conclude that $C' J(1, \mathbf{\Phi}_{\ell b}^L) ||\overline{M}||_{P,2} \leq C' J(1, \mathbf{\Phi}_{\ell b}^L) C^{2/(2+\delta)} < \infty$. Combining this with (S.2.16), (S.2.17) and (S.2.19) completes the proof. ■

S.3 Additional Comparison to Christensen and Connault (2023)

In this section we revisit Example 1 and Example 2, and compare our method with the method of Christensen and Connault (2023). These examples are chosen because they were also used as examples in Section 2.1 of Christensen and Connault (2023). We then comment on the connection between our approach and the nonparametric bounds considered in Section 2.5 of Christensen and Connault (2023).

S.3.1 Comparison in Specific Examples

Example 1 (Multinomial Choice (Cont'd))**.** *Consider again the example of multinomial choice in the main text, where consumers select among J alternatives, and the utility obtained from choice $j \in \mathcal{J} = \{1, \ldots, J\}$ is:*

$$\pi_j(\boldsymbol{z}_i, \boldsymbol{u}_i, \theta) = \boldsymbol{z}_{ij}^\top \theta_j + u_{ij},$$

where $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ})$ takes values in a finite set $\mathcal{Z}$, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iJ})$, and $\theta = (\theta_1, \ldots, \theta_J)$. Let $Y_i \in \mathcal{J}$ denote the choice of consumer i, and assume the latent variables $\boldsymbol{U}_i$ lie in a $1-$Wasserstein ball of radius ρ around $\boldsymbol{U}_i'$, which has a known distribution $P_{\boldsymbol{U}'}$. Now consider the problem of

bounding the social surplus:

$$W(\boldsymbol{z}) := \mathbb{E}\left[\max_{j\in\mathcal{J}}\{\boldsymbol{z}_j^\top\theta_j + U_j\}\right],$$

where $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_J)$ *is a fixed vector. Here the counterfactual function is given by:*

$$\varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) = \varphi(\boldsymbol{u}, \theta) = \max_{j\in\mathcal{J}}\{\boldsymbol{z}_j^\top\theta_j + u_j\}.$$

By Theorem 2.3, the lower bound for a fixed (θ, τ) *is given by:*

$$\max_{\mu\in\mathbb{R}_+}\int\left[\inf_{\boldsymbol{u}\in\mathcal{U}(Y,\boldsymbol{Z},\theta)}\varphi(\boldsymbol{u},\theta) + \mu\left(\sum_{j=1}^{J}|u_j - U_j'| - \rho + \tau\right)\right]d\mathbb{P}. \tag{S.3.1}$$

where:

$$\mathcal{U}(y, \boldsymbol{z}, \theta) = \left\{\boldsymbol{u}\in\mathbb{R}^J : \boldsymbol{z}_y^\top\theta_y + u_y \geq \boldsymbol{z}_{y'}^\top\theta_j + u_{y'}, \forall y' \neq y\right\}.$$

With a finite sample, estimation proceeds by replacing the expectation in (S.3.1) with its sample analog. Following a similar derivation as in Section 3.1, the local problem for each observation can be written as a linear program. The middle problem is only a $1-$*dimensional convex optimization problem, and following Remark 3.1 there are at most* $|\mathcal{Y}| \times |\mathcal{Z}|$ *linear programs to solve at each iteration.*

Assume now that $P_{\boldsymbol{U}}$ *lies in a Kullback-Liebler neighborhood* $\mathcal{N}_\rho$ *of a baseline distribution* $P_{\boldsymbol{U}'}$*, and consider the approach taken in Section 2.1 of Christensen and Connault (2023). In this model, only the moment functions* $\boldsymbol{m}_2$ *from (2.16) are needed. In this case, set:*

$$\boldsymbol{m}_2(\boldsymbol{u}, \theta, \gamma) = \boldsymbol{m}_2(\boldsymbol{u}, \theta) = \left(\mathbb{1}\left\{\boldsymbol{z}_y^\top\theta_y + U_y = \max_{y'\in\mathcal{J}}\boldsymbol{z}_{y'}^\top\theta_{y'} + U_{y'}\right\}\right)_{(y,\boldsymbol{z})\in\mathcal{Y}\times\mathcal{Z}},$$

and set $\boldsymbol{q}_{20} = (\mathbb{P}(Y = y \mid \boldsymbol{Z} = \boldsymbol{z}))_{(y,\boldsymbol{z})\in\mathcal{Y}\times\mathcal{Z}}$*, the true vector of conditional choice probabilities.*[2] *By Proposition 2.1 in Christensen and Connault (2023), the lower bound is determined by the program:*

$$\sup_{\eta>0,\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{Y}\times\mathcal{Z}|}} -\eta\log\left(\int\exp\left(\frac{\varphi(\boldsymbol{U}', \theta) + \boldsymbol{\lambda}^\top\boldsymbol{m}_2(\boldsymbol{U}', \theta)}{-\eta}\right)d\mathbb{P}\right) - \eta\rho - \boldsymbol{\lambda}^\top\boldsymbol{q}_{20}. \tag{S.3.2}$$

Compared to (S.3.1), there is no local optimization problem to solve. The optimization problem over $\eta > 0$ *and* $\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{Y}\times\mathcal{Z}|}$ *is of higher dimension than the optimization problem over* μ *in (S.3.1), but unlike in (S.3.1) the objective function in (S.3.2) is differentiable in* $(\eta, \boldsymbol{\lambda})$*, which greatly aids with computation.*

Example 2 (Binary Game (Cont'd))**.** *Consider again the binary game with complete information,*

[2] See Example 2.1 in Christensen and Connault (2023).

pure strategy Nash equilibria, and $K = 2$ players, where the observed entry decisions must satisfy:

$$y_{i1} = \mathbb{1}\{\boldsymbol{z}_{i1}^\top \beta_1 + y_{i2}\delta_1 \geq u_{i1}\}, \qquad y_{i2} = \mathbb{1}\{\boldsymbol{z}_{i2}^\top \beta_2 + y_{i1}\delta_2 \geq u_{i2}\}.$$

Suppose that the latent variables $\boldsymbol{U}_i$ are assumed to lie in a $1-$Wasserstein ball of radius ρ around $\boldsymbol{U}_i'$, which has a known distribution $P_{\boldsymbol{U}'}$. For the sake of comparison, suppose that $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ takes values in a finite set $\mathcal{Z}$. Now consider a counterfactual that replaces $(\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ with $(\check{\boldsymbol{z}}_{i1}, \check{\boldsymbol{z}}_{i2})$, and simultaneously forces $y_{i2}^\star = 0$. Suppose the researcher wishes to study the effect of this intervention on player 1's entry probability. The counterfactual outcomes $(y_{i1}^\star, y_{i2}^\star)$ must satisfy:

$$y_{i1}^\star = \mathbb{1}\{\check{\boldsymbol{z}}_{i1}^\top \beta_1 + y_{i2}^\star \delta_1 \geq u_{i1}\}, \qquad y_{i2}^\star = 0.$$

To bound player 1's counterfactual entry probability, set:

$$\varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) = \varphi(\boldsymbol{u}, \theta) = \mathbb{1}\{\check{\boldsymbol{z}}_{i1}^\top \beta_1 \geq u_{i1}\}.$$

By Theorem 2.3, the lower bound for a fixed (θ, τ) is given by:

$$\max_{\mu \in \mathbb{R}_+} \int \left[\inf_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)} \varphi(\boldsymbol{u}, \theta) + \mu \left(\sum_{k=1}^{2} |u_j - U_j'| - \rho + \tau \right) \right] d\mathbb{P}. \tag{S.3.3}$$

where:

$$\mathcal{U}(\boldsymbol{y}, \boldsymbol{z}, \theta) = \left\{ \boldsymbol{u} \in \mathbb{R}^2 : \ y_1 = \mathbb{1}\{\boldsymbol{z}_1^\top \beta_1 + y_2 \delta_1 \geq u_1\}, \ y_2 = \mathbb{1}\{\boldsymbol{z}_2^\top \beta_2 + y_1 \delta_2 \geq u_2\} \right\}.$$

With a finite sample, estimation proceeds by replacing the expectation in (S.3.3) with its sample analog. Following a similar derivation as in Section 3.1, the local problem for each observation can be written as a mixed integer linear program with one integer variable, two continuous variables, and 6 linear constraints. The middle problem is a $1-$dimensional convex optimization problem, and following Remark 3.1 there are at most $|\mathcal{Y}| \times |\mathcal{Z}|$ mixed integer linear programs to solve at each iteration in the middle problem.

Suppose now that $P_{\boldsymbol{U}}$ belongs to a KL neighborhood $\mathcal{N}_\rho$ of a baseline distribution $P_{\boldsymbol{U}'}$, and consider the approach taken in Section 2.1 of Christensen and Connault (2023). Following Example 2.2 in Christensen and Connault (2023), set:

$$\boldsymbol{m}_1(\boldsymbol{u}, \theta, \gamma) = \boldsymbol{m}_1(\boldsymbol{u}, \theta) = \begin{bmatrix} \left(-\mathbb{1}\left\{\boldsymbol{z}_1^\top \beta_1 \geq u_1\right\} \mathbb{1}\left\{\boldsymbol{z}_2^\top \beta_2 + \delta_2 < u_2\right\}\right)_{\boldsymbol{z} \in \mathcal{Z}} \\ \left(-\mathbb{1}\left\{\boldsymbol{z}_1^\top \beta_1 + \delta_1 < u_1\right\} \mathbb{1}\left\{\boldsymbol{z}_2^\top \beta_2 \geq u_2\right\}\right)_{\boldsymbol{z} \in \mathcal{Z}} \end{bmatrix},$$

$$\boldsymbol{m}_2(\boldsymbol{u}, \theta, \gamma) = \boldsymbol{m}_2(\boldsymbol{u}, \theta) = \begin{bmatrix} \left(\mathbb{1}\left\{\boldsymbol{z}_1^\top \beta_1 < u_1\right\} \mathbb{1}\left\{\boldsymbol{z}_2^\top \beta_2 < u_2\right\}\right)_{\boldsymbol{z} \in \mathcal{Z}} \\ \left(\mathbb{1}\left\{\boldsymbol{z}_1^\top \beta_1 + \delta_1 \geq u_1\right\} \mathbb{1}\left\{\boldsymbol{z}_2^\top \beta_2 + \delta_2 \geq u_2\right\}\right)_{\boldsymbol{z} \in \mathcal{Z}} \end{bmatrix},$$

and:

$$\boldsymbol{q}_{10} = \begin{bmatrix} -\left(\mathbb{P}((Y_1, Y_2) = (1,0) \mid \boldsymbol{Z} = \boldsymbol{z})\right)_{\boldsymbol{z} \in \mathcal{Z}} \\ -\left(\mathbb{P}((Y_1, Y_2) = (0,1) \mid \boldsymbol{Z} = \boldsymbol{z})\right)_{\boldsymbol{z} \in \mathcal{Z}} \end{bmatrix}, \quad \boldsymbol{q}_{20} = \begin{bmatrix} -\left(\mathbb{P}((Y_1, Y_2) = (0,0) \mid \boldsymbol{Z} = \boldsymbol{z})\right)_{\boldsymbol{z} \in \mathcal{Z}} \\ -\left(\mathbb{P}((Y_1, Y_2) = (1,1) \mid \boldsymbol{Z} = \boldsymbol{z})\right)_{\boldsymbol{z} \in \mathcal{Z}} \end{bmatrix}.$$

By Proposition 2.1 in Christensen and Connault (2023), the lower bound is determined by the program:

$$\sup_{\eta > 0, \boldsymbol{\lambda}_1 \in \mathbb{R}_+^{2|\mathcal{Z}|}, \boldsymbol{\lambda}_2 \in \mathbb{R}^{2|\mathcal{Z}|}} -\eta \log \left(\int \exp \left(\frac{\varphi(\boldsymbol{U}', \theta) + \boldsymbol{\lambda}_1^\top \boldsymbol{m}_1(\boldsymbol{U}', \theta) + \boldsymbol{\lambda}_2^\top \boldsymbol{m}_2(\boldsymbol{U}', \theta)}{-\eta} \right) d\mathbb{P} \right) \tag{S.3.4}$$
$$- \eta\rho - \boldsymbol{\lambda}_1^\top \boldsymbol{q}_{10} - \boldsymbol{\lambda}_2^\top \boldsymbol{q}_{20}.$$

Compared to (S.3.3), there is no local optimization problem to solve. However, the optimization problem over $\eta > 0$ and $\boldsymbol{\lambda}_1 \in \mathbb{R}_+^{2|\mathcal{Z}|}$ and $\boldsymbol{\lambda}_2 \in \mathbb{R}^{2|\mathcal{Z}|}$ is of higher dimension than the optimization problem over μ in (S.3.1), which is always one dimensional. Furthermore, the dimension of the optimization problem in (S.3.4) scales linearly in $|\mathcal{Z}|$ and exponentially in $|\mathcal{Y}|$ (e.g. the number of players). The moment conditions used in (S.3.4) are not sharp, although in the two player example this can be amended by including an additional set of $2|\mathcal{Z}|$ moment inequalities, corresponding to the insights of Ciliberto and Tamer (2009). There are alternative sharp characterizations of the identified set for the game example in the case when $K > 2$ (e.g. see the discussion of Artstein's inequalities in Beresteanu, Molchanov, and Molinari (2011)), although the required number of inequalities can quickly become intractable. Furthermore, including them in (S.3.4) requires enumerating all equilibria for each draw of $\boldsymbol{U}'$, which can become a bottleneck with many players. Finally, incomplete counterfactuals (e.g. when player 2 can respond endogenously) can be accommodated in our framework by incorporating the set $\mathcal{Y}^\star(\boldsymbol{z}, \boldsymbol{u}, \theta)$ into the problem (S.3.3) using an additional set of constraints (see Section 3.1 of the main text). In contrast, sharp bounds on counterfactuals in incomplete counterfactuals are more challenging to accommodate in the framework of Christensen and Connault (2023), since the counterfactual functional is not uniquely determined by the value of the latent variables.

S.3.2 Further Connections

As highlighted in Section 2.4 and the previous section, there are a number of differences between our framework and the framework of Christensen and Connault (2023). However, there are some connections between our approach and the nonparametric bounds of Christensen and Connault (2023). Here we give an informal sketch of an argument relating the two bounds. First, suppose

that there exists a vector of moment functions $\boldsymbol{m}$ such that:[3]

$$\exists(\theta, \boldsymbol{U}) \text{ s.t. } \boldsymbol{U} \in \mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta) \iff \exists(\theta, \boldsymbol{U}) \text{ s.t. } \mathbb{E}[\boldsymbol{m}(\boldsymbol{U}, \theta)] = \boldsymbol{q}, \tag{S.3.5}$$

where $\boldsymbol{q}$ is the form $\boldsymbol{q} = \mathbb{E}[\boldsymbol{h}(\boldsymbol{Y}, \boldsymbol{Z})]$ for some measurable vector-valued function $\boldsymbol{h} : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^{d_m}$. Now consider a complete counterfactual model, and suppose the researcher is interested in the counterfactual quantity $\mathbb{E}[\varphi(\boldsymbol{U}, \theta)]$. Under some assumptions Christensen and Connault (2023) show that the nonparametric lower bound on $\mathbb{E}[\varphi(\boldsymbol{U}, \theta)]$ (i.e. taking the neighborhood radius ρ around some baseline $P_{\boldsymbol{U}'}$ to ∞) is given by:

$$\underline{K}_{\infty}(\theta, \boldsymbol{q}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m} : \text{ ess inf}(\varphi(\,\cdot\,) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\,\cdot\,, \theta)) > -\infty} \text{ess inf}(\varphi(\,\cdot\,) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\,\cdot\,, \theta) - \boldsymbol{\lambda}^\top \boldsymbol{q}),$$

where the essential infimum is with respect to the baseline measure $P_{\boldsymbol{U}'}$. When $P_{\boldsymbol{U}'}$ and the Lebesgue measure are mutually absolutely continuous, this essential infimum can be taken instead with respect the Lebesgue measure. In addition, if there exists at least one $\boldsymbol{\lambda} \in \mathbb{R}^{d_m}$ such that $\text{ess inf}(\varphi(\,\cdot\,) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\,\cdot\,, \theta) - \boldsymbol{q} \rangle) > -\infty$, then the lower bound can be written as:

$$\underline{K}_{\infty}(\theta, \boldsymbol{q}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \text{ess inf}(\varphi(\,\cdot\,) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\,\cdot\,, \theta) - \boldsymbol{\lambda}^\top \boldsymbol{q}).$$

Finally, if $\varphi(\,\cdot\,)$ and $\boldsymbol{m}(\,\cdot\,)$ are continuous (for example), then the Lebesgue essential infimum is equal to the infimum, in which case the previous display becomes:

$$\underline{K}_{\infty}(\theta, \boldsymbol{q}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \inf_{\boldsymbol{u} \in \mathcal{U}} \varphi(\boldsymbol{u}) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\boldsymbol{u}, \theta) - \boldsymbol{\lambda}^\top \boldsymbol{q}. \tag{S.3.6}$$

Now consider the lower bound constructed using Theorem 2.1. Under condition (S.3.5), all restrictions on the latent variables $\boldsymbol{U}$ can be expressed using moment conditions, in which case the $\boldsymbol{U}-$level set $\mathcal{U}(\boldsymbol{Y}, \boldsymbol{Z}, \theta)$ can be replaced with $\mathcal{U}$. In this case the lower bound is:

$$\begin{aligned}
&\sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \int \left[\inf_{\boldsymbol{u} \in \mathcal{U}} \varphi(\boldsymbol{u}) + \boldsymbol{\lambda}^\top (\boldsymbol{m}(\boldsymbol{u}, \theta) - \boldsymbol{h}(\boldsymbol{Y}, \boldsymbol{Z})) \right] d\mathbb{P} \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \inf_{\boldsymbol{u} \in \mathcal{U}} \varphi(\boldsymbol{u}) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\boldsymbol{u}, \theta) - \boldsymbol{\lambda}^\top \int \boldsymbol{h}(\boldsymbol{Y}, \boldsymbol{Z}) \, d\mathbb{P} \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{d_m}} \inf_{\boldsymbol{u} \in \mathcal{U}} \varphi(\boldsymbol{u}) + \boldsymbol{\lambda}^\top \boldsymbol{m}(\boldsymbol{u}, \theta) - \boldsymbol{\lambda}^\top \boldsymbol{q},
\end{aligned}$$

matching the lower bound from (S.3.6). This informal derivation therefore shows that the bounds are identical in certain cases.

[3] Here we focus on moment equalities for simplicity. The same argument follows if moment inequalities are included.

S.4 Additional Algorithm Details

Here we provide some additional details on the algorithm, with a particular focus on the settings used for the application, which bounds a counterfactual entry probability.

S.4.1 The Local Problem

In the application, the mixed integer linear programs are implemented in R using Mosek. As mentioned in Example 2, solving the mixed integer linear programs requires a choice of the parameters M and η in order to approximate the indicator constraints. The approximation error is smaller when M is larger and η is smaller. However, values of M that are too large or values of η that are too small can cause numerical instabilities. We recommend $M = 10^4$ and $\eta = 10^{-4}$, which were selected after some calibration. Because our choice of η limits precision in the local problem, we rounded the draws $\{\boldsymbol{u}_i'\}_{i=1}^n$ to the third decimal place, and also rounded the optimal $\boldsymbol{u}_i^*$'s in the local problem to the third decimal place before computing the Wasserstein penalty.

S.4.2 The Middle Problem

To implement the Wasserstein penalty, we always initially draw $\{\boldsymbol{u}_i'\}_{i=1}^n$ using quasi-random Halton sequences. After doing so, the middle problem begins by evaluating the average local problem at $\mu = 10^3$ to check if the lower bound is above 1, or if the upper bound is below 0 (since we're bounding an entry probability). At this step, the matching procedure described in Section 3.3 is run to full convergence. If the lower bound is above 1, or the upper bound is below 0, this is sufficient to rule out that the current θ belongs to the identified set, and we move to the next value of θ.[4] If the initial $\mu = 10^3$ returns an average local problem between 0 and 1, we then evaluate the average local problem on the coarse grid $\mu \in \{1, 2, 4, 8, 16\}$. For each of these values, we also run the procedure from Section 3.3 to full convergence, and save the final values of $\{\boldsymbol{u}_i'\}_{i=1}^n$. We then choose the value of $\mu \in \{1, 2, 4, 8, 16\}$ with the largest (or smallest) value of the middle problem for the lower (upper) bound, and start stochastic subgradient descent (SSGD) from that value, initializing with the corresponding $\{\boldsymbol{u}_i'\}_{i=1}^n$.

Algorithm 1 shows the four SSGD algorithms considered in the application section. The step size plays an important role in SSGD algorithms, and all four of the algorithms have adaptive stepsizes. However, most of the algorithms depend on a user-chosen learning rate η. AdaGrad is a quasi-Newton method that approximates the Hessian with the root of a diagonal matrix $\boldsymbol{G}_k$, where the main diagonal is equal to the sum of the squares of all previous gradients (see Algorithm 1). However, the diagonal of the matrix $\boldsymbol{G}_k$ is monotonically increasing, and the AdaGrad step size often decreases too quickly. Both RMSprop and AdaDelta guard against this by replacing $\boldsymbol{G}_k$ with

[4] In this sense, boundedness of our counterfactual functional is helpful.

Algorithm 1: (SSGD, General)

input θ, $\boldsymbol{\lambda}_0$, K, η (AdaGrad, RMSprop), γ_1 (RMSprop, AdaDelta, Adam), γ_2 (Adam)
for $k = 1, \ldots, K$**:**

Sample i_k uniformly from $\{1, \ldots, n\}$.

Solve the local problem at i_k, let $(\boldsymbol{u}^*_{i_k}, \boldsymbol{y}^{**}_{i_k})$ denote any optimal solution.

Set $\boldsymbol{g}_{i_k} := \boldsymbol{m}(\boldsymbol{y}_{i_k}, \boldsymbol{z}_{i_k}, \boldsymbol{u}^*_{i_k}, \theta)$.

Update:

AdaGrad: $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \eta \boldsymbol{G}_k^{-1/2} \boldsymbol{g}_{i_k}$, where $\boldsymbol{G}_k = \text{diag}(\sum_{t=1}^k \boldsymbol{g}_{i_t} \boldsymbol{g}_{i_t}^\top)$.

RMSprop: $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \eta \boldsymbol{G}_k^{-1/2} \boldsymbol{g}_{i_k}$, where $\boldsymbol{G}_0 = \boldsymbol{0}$ and:

$$\boldsymbol{G}_k = \gamma_1 \boldsymbol{G}_{k-1} + (1-\gamma_1) diag(\boldsymbol{g}_{i_k} \odot \boldsymbol{g}_{i_k}), \quad \forall k \geq 1.$$

AdaDelta: $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \boldsymbol{E}_{k-1}^{1/2} (\boldsymbol{G}_k)^{-1/2} \boldsymbol{g}_{i_k}$, where $\boldsymbol{G}_0 = \boldsymbol{0}$, $\boldsymbol{E}_0 = \boldsymbol{0}$, and:

$$\begin{aligned} \boldsymbol{G}_k &= \gamma_1 \boldsymbol{G}_{k-1} + (1-\gamma_1) diag(\boldsymbol{g}_{i_k} \odot \boldsymbol{g}_{i_k}), \quad \forall k \geq 1, \\ \boldsymbol{E}_{k-1} &= \gamma_1 \boldsymbol{E}_{k-2} + (1-\gamma_1) diag((\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k-1}) \odot (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k-1})), \quad \forall k \geq 1. \end{aligned}$$

Adam: $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - (1-\gamma_1^k)/(1-\gamma_2^k) \boldsymbol{F}_k (\boldsymbol{G}_k)^{-1/2}$, where:

$$\begin{aligned} \boldsymbol{G}_k &= \gamma_1 \boldsymbol{G}_{k-1} + (1-\gamma_1) diag(\boldsymbol{g}_{i_k} \odot \boldsymbol{g}_{i_k}), \quad \forall k \geq 1, \\ \boldsymbol{F}_k &= \gamma_2 \boldsymbol{F}_{k-1} + (1-\gamma_2) \boldsymbol{g}_{i_k}, \quad \forall k \geq 1. \end{aligned}$$

end
return $(\boldsymbol{\lambda}_K, \hat{\varphi}^L_{\ell b}(\boldsymbol{\lambda}_K, \theta))$.

an exponentially decaying weighted average of the squared gradients. AdaDelta also replaces the learning rate parameter η with an exponentially decaying weighted average of the previous squared step sizes, and so has a fully adaptive learning rate. Finally, the Adam algorithm approximates the mean and variance of the gradients using exponentially decaying averages, and also includes a bias correction term to correct for a bias towards small steps sizes that occurs early in most SSGD procedures (see Algorithm 1). Note the Adam updates nest both the AdaGrad and RMSprop updates as special cases. We revisit these algorithms in the bench-marking exercise in Section 5.

Although they perform well in our setting, the SSGD algorithms developed in the machine learning literature are designed for problems with different structure. For instance, the parameter $\boldsymbol{\lambda}$ is often high-dimensional in machine learning settings, and the gradients can be sparse. This has in turn influenced the design of the algorithms, which attempt to balance computational efficiency with accuracy, and typically avoid operations that scale poorly with the dimension of $\boldsymbol{\lambda}$. Our problem is also slightly different since we expect the middle problem to converge only for $\theta \in \Theta^*$, and expect it to diverge otherwise. For this to occur, the step size should tend to zero as fast as possible for $\theta \in \Theta^*$, but should ideally tend to $+\infty$ for $\theta \notin \Theta^*$. The adaptive step sizes of the SSGD updates in Algorithm 1 work well for our purposes, but there may be considerable opportunities to improve the SSGD updates in Algorithm 1.

Algorithm	Tuning Parameters
AdaGrad	$\eta = 0.01$
RMSprop	$\eta = 0.001$, $\gamma = 0.9$
AdaDelta	–
Adam	$\eta = 0.001$, $\gamma_1 = 0.9$, $\gamma_2 = 0.99$

Table 1: The tuning parameters for the SSGD algorithms presented in Section 3.2.

The tuning parameters for each of the algorithms from Section 3.2 are provided in Table 1.[5] The SSGD procedure is run in *epochs*, with one epoch representing a full pass through the sample. Every two epochs, we compute the value of the average local problem on the full sample to check convergence. At this point we also update the values of $\{\boldsymbol{u}_i'\}_{i=1}^n$ using the procedure from Section 3.3. Convergence is achieved if the problem makes no progress for ten epochs. In particular, if the change in the objective function is less than 10^{-4} between two epochs, and if this happens five consecutive times (i.e. ten consecutive epochs), then the middle problem terminates.

S.4.3 The Outer Problem

The outer problem begins by randomly drawing $10d_\theta+1$ initial evaluation points from the parameter space using latin hypercube sampling. In the application, the parameter space (before imposing sign restrictions) is taken as $\Theta = [-1, 1]^{d_\theta}$, which was selected after some initial exploration of the parameter space. After the initial evaluation of the $10d_\theta + 1$ points, we fit a Gaussian process regression model using the `GPfit` package in R. The hyperparameters μ, σ^2, and the parameters in the covariance kernel do not change substantially as new additional evaluation points are added, so we update these hyperparameters only every ten iterations. Maximizing expected improvement at each iteration is delicate, and most off-the-shelf optimizers perform poorly at this step. Of all the optimizers we tested, differential evolution and particle swarm methods perform the best, and give consistent results. Differential evolution was also recommended in a different context by Beresteanu, Molchanov, and Molinari (2011), and we use the R package `DEOptim`. The algorithm terminates when expected improvement has fallen by a factor of 10^3 from the first evaluation of expected improvement.

S.4.4 Post-Processing

After obtaining the initial output, we also perform some post-processing to improve the quality of the solutions. In particular, for each of the lower bounds, we take the four values of θ that have the smallest bounds from the initial output, and re-run SSGD until convergence using the full sample at each iteration (rather than a single observation) while also running the matching procedure described in Section 3.3 after each iteration. The full Adam iterations are initialized

[5] See Ruder (2016) for a discussion.

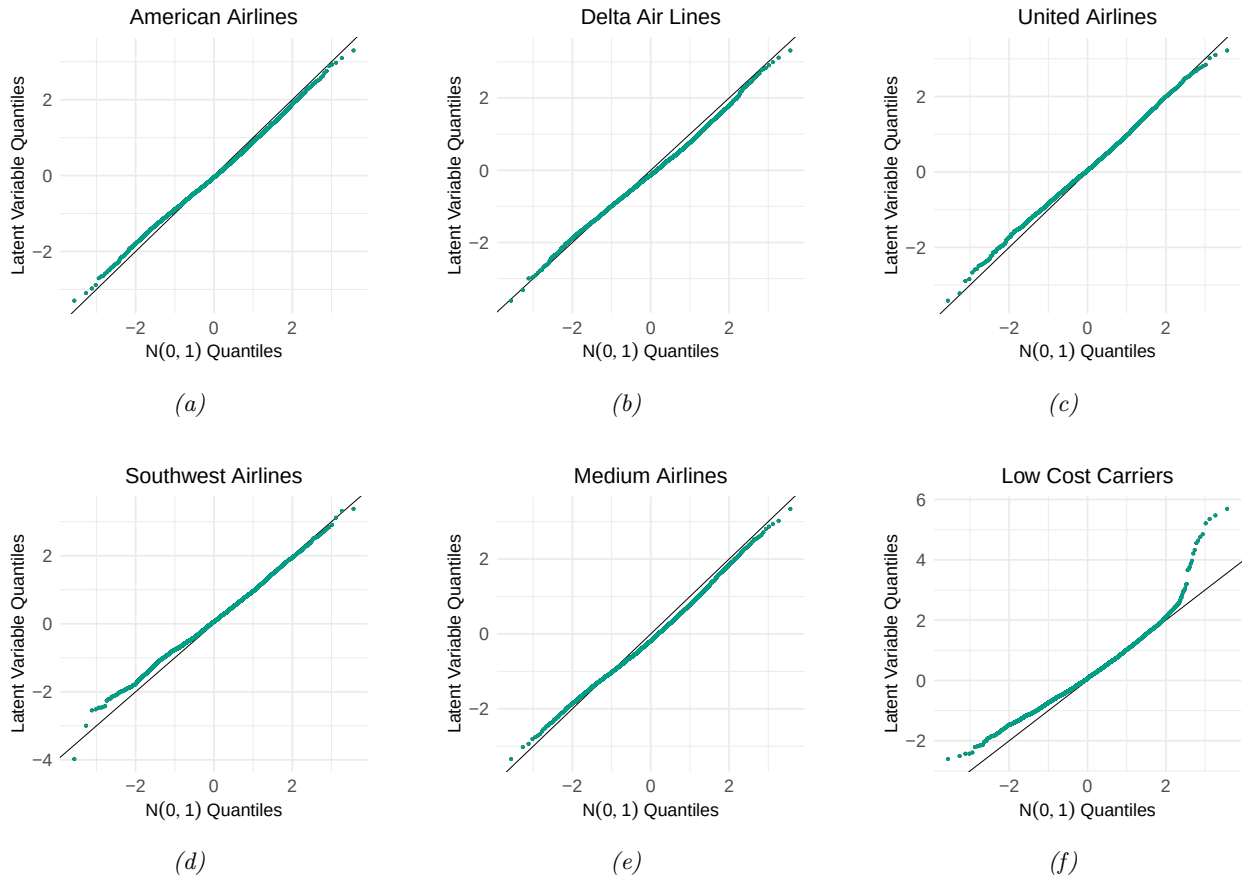


Figure 1: Quantile-quantile plots for the latent variables of all six players. Here $\rho = 0.84$ in the six-player game when imposing independence between the payoff-relevant latent variables in the baseline distribution.

using the results from the SSGD algorithm from the initial run. This procedure is expensive, but converges quickly in most cases. Although rare, when this procedure produces substantially different values of the lower bound, we continue applying the procedure to other values of θ from the initial output until the lower bound stabilizes. A similar procedure was repeated for the upper bound. We found this procedure improves the quality of the solutions substantially, and would recommend it to researchers interested in the method.

As a final post-processing step, we also take the values of θ that obtain the lower (or upper) bound for one value of ρ, and try them at neighboring values of ρ. This "cross-pollination" of the values of θ is a final way to check that each bounding problem does not miss potentially promising values of θ found by neighboring problems.

S.5 Additional Results

S.5.1 Additional Application Results

Figure 1 shows quantile-quantile plots comparing the quantiles of U_{ik}, obtained from the solution to the local problems when $\boldsymbol{U}_i' \sim N(\mathbf{0}, \boldsymbol{I})$, to the quantiles of a standard normal for all six players. These plots are obtained when maximizing the counterfactual entry probability of American Airlines with $\rho = 0.84$ subject to a $1-$Wasserstein constraint. Note that for nearly all players the latent variable distributions reconstructed from the solutions to the local problem have quantiles that are very close to the baseline standard normal distribution. The exception is the low cost carriers, which show substantial deviations from the standard normal at the upper quantiles.

S.5.2 Verifying Dudley's Entropy Condition

We will focus on the class of functions:

$$\boldsymbol{\Phi}_{\ell b}^{L} := \left\{ \varphi_{\ell b}^{L}(\,\cdot\,, \theta, \boldsymbol{\lambda}) : (\theta, \boldsymbol{\lambda}) \in \Theta \times \Lambda \right\},$$

equipped with the envelope function $\overline{M} : \mathcal{Y} \times \mathcal{Z} \to [0, \infty)$. Finally, let us define:

$$\begin{aligned} \Phi &:= \{\varphi(\boldsymbol{y}^{\star}, \,\cdot\,, \boldsymbol{u}, \theta) : \mathcal{Z} \to \mathbb{R} \,|\, (\boldsymbol{u}, \boldsymbol{y}^{\star}) \in \mathcal{U} \times \mathcal{Y}\}, \\ \mathcal{M}_j &:= \{m_j(\,\cdot\,, \boldsymbol{u}, \theta) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} \,|\, (\boldsymbol{u}, \theta) \in \mathcal{U} \times \Theta\}, \;\; j = 1, \ldots, d_m. \end{aligned}$$

Note that these function classes treat both $\boldsymbol{u} \in \mathcal{U}$ and $\boldsymbol{y}^{\star} \in \mathcal{Y}$ as parameters. The following is Definition 2.1 in Chernozhukov, Chetverikov, and Kato (2014).

Definition S.5.1 (VC-Type Class)**.** *Let $\mathcal{F}$ be a class of measurable real-valued functions on a measurable space $(\mathcal{X}, \mathscr{A})$ with measurable envelope F. We say $\mathcal{F}$ is a VC-type class of functions with envelope F if there exists constants A and $v > 0$ such that:*

$$\sup_{Q} N\left(\epsilon \cdot ||F||_{Q,2}, \mathcal{F}, ||\,\cdot\,||_{Q,2}\right) \precsim \left(\frac{A}{\epsilon}\right)^{v},$$

for all $0 < \epsilon \leq 1$, where the supremum is taken over all finitely discrete probability measures Q on $(\mathcal{X}, \mathscr{A})$ and where $\precsim$ denotes an inequality up to multiplication by a constant.

Importantly, VC-type classes can have unbounded envelope functions. Furthermore, every VC-subgraph class is VC-type, although VC-type classes include functions classes that are not VC-subgraph. The following result implies that, if $\Phi, \mathcal{M}_1, \ldots, \mathcal{M}_{d_m}$ are VC-type, then Dudley's entropy condition holds for the class $\boldsymbol{\Phi}_{\ell b}^{L}$.

Lemma S.5.1. *Suppose Assumptions 2.1, 2.2, and 4.1 hold. Furthermore, suppose that* $\Phi, \mathcal{M}_1, \ldots, \mathcal{M}_{d_m}$ *are VC-type with envelope functions* $\bar{\Phi}, \overline{M}_1, \ldots, \overline{M}_{d_m}$*, respectively. Then:*

$$\sup_{Q \in \mathcal{Q}} \log N\left(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, ||\cdot||_{Q,2}\right) \precsim \left(\frac{1}{\epsilon}\right)^{\nu},$$

for some $0 \leq \nu < 2$ *and every* $0 < \epsilon \leq 1$.

Proof of Lemma S.5.1. Define the following functions:

$$\begin{aligned} f_I(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{y}^\star, \theta, \boldsymbol{\lambda}) &:= \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) + \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta)\rangle, \\ f_{II}(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta) &:= \varphi(\boldsymbol{y}^\star, \boldsymbol{z}, \boldsymbol{u}, \theta), \\ f_{III}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta, \boldsymbol{\lambda}) &:= \langle \boldsymbol{\lambda}, \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \theta)\rangle. \end{aligned}$$

In addition, define the corresponding classes:

$$\begin{aligned} \mathcal{F}_I &:= \{f_I(\,\cdot\,, \boldsymbol{u}, \boldsymbol{y}^\star, \theta, \boldsymbol{\lambda}) : (\boldsymbol{u}, \boldsymbol{y}^\star, \theta, \boldsymbol{\lambda}) \in \mathcal{U} \times \mathcal{Y} \times \Theta \times \Lambda\}, \\ \mathcal{F}_{II} &:= \{f_{II}(\,\cdot\,, \boldsymbol{u}, \boldsymbol{y}^\star, \theta) : (\boldsymbol{u}, \boldsymbol{y}^\star) \in \mathcal{U} \times \mathcal{Y} \times \Theta\}, \\ \mathcal{F}_{III} &:= \{f_{III}(\,\cdot\,, \boldsymbol{u}, \theta, \boldsymbol{\lambda}) : (\boldsymbol{u}, \theta, \boldsymbol{\lambda}) \in \mathcal{U} \times \Theta \times \Lambda\}. \end{aligned}$$

Without loss of generality, set $\overline{L} := d_m \left(\sup_{\boldsymbol{\lambda} \in \Lambda} \max_{k=1,\ldots,d_m} |\lambda_k M_k|\right)$, and take $\overline{M} \geq \overline{\Phi} + \overline{L}$. By Lemma S.5.3, we have:

$$N\left(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, ||\cdot||_{Q,2}\right) \leq N\left((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_I, ||\cdot||_{Q,2}\right).$$

By Lemma S.5.4 we also have:

$$\begin{aligned} &N((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_I, ||\cdot||_{Q,2}) \\ &\qquad \leq N((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_{II}, ||\cdot||_{Q,2}) N((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_{III}, ||\cdot||_{Q,2}). \end{aligned}$$

Conclude that:

$$\begin{aligned} &\sup_{Q \in \mathcal{Q}} \log N\left(\epsilon \cdot ||\overline{M}||_{Q,2}, \mathbf{\Phi}_{\ell b}^L, ||\cdot||_{Q,2}\right) \\ &\leq \sup_{Q \in \mathcal{Q}} \log N((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_{II}, ||\cdot||_{Q,2}) + \sup_{Q \in \mathcal{Q}} \log N((\epsilon/2) \cdot ||\overline{M}||_{Q,2}, \mathcal{F}_{III}, ||\cdot||_{Q,2}) \\ &\leq \sup_{Q \in \mathcal{Q}} \log N((\epsilon/2) \cdot ||\overline{\Phi}||_{Q,2}, \mathcal{F}_{II}, ||\cdot||_{Q,2}) + \sup_{Q \in \mathcal{Q}} \log N((\epsilon/2) \cdot ||\overline{L}||_{Q,2}, \mathcal{F}_{III}, ||\cdot||_{Q,2}). \end{aligned}$$

Since $\mathcal{M}_1, \ldots, \mathcal{M}_J$ are VC-type with square-integrable envelope functions, Lemma S.5.2 implies:

$$\sup_{Q \in \mathcal{Q}} \log N((\epsilon/2) \cdot ||\overline{L}||_{Q,2}, \mathcal{F}_{III}, ||\cdot||_{Q,2}) \precsim \left(\frac{1}{\epsilon}\right)^{v},$$

for every $0 < \epsilon \leq 1$ for some $0 \leq v < 2$. The proof then follows from the fact that Φ is VC-type

with square integrable envelope $\overline{\Phi}$.

■

Lemma S.5.2. *Let Q be any probability measure on the measurable space $(\mathcal{X}, \mathscr{A})$, let $\mathcal{F}_1, \ldots, \mathcal{F}_K$ be VC-type classes of real-valued measurable functions on $\mathcal{X}$ with measurable envelope functions $F_1, \ldots, F_K$. For $B \geq 1$ consider the class of functions:*

$$\mathcal{G}_B := \left\{ \sum_{k=1}^{K} t_k f_k(\boldsymbol{x}) : \sum_{k=1}^{K} t_k \leq B \text{ and } f_k \in \mathcal{F}_k \right\}.$$

Then there exists a value $0 \leq v < 2$ such that:

$$\log N(\epsilon ||G||_{Q,2}, \mathcal{G}_B, ||\cdot||_{Q,2}) \precsim \left(\frac{1}{\epsilon}\right)^v,$$

where $G(\boldsymbol{x}) \geq B||\boldsymbol{F}(\boldsymbol{x})||_1$ and $\boldsymbol{F}(\boldsymbol{x}) := (F_1(\boldsymbol{x}), \ldots, F_K(\boldsymbol{x}))$.

Proof of Lemma S.5.2. Define the class $\mathcal{F} := \mathcal{F}_1 \cup \ldots \cup \mathcal{F}_K$, and let $B\mathcal{F} = \{Bf : f \in \mathcal{F}\}$. For any function $f \in B\mathcal{F}$ we have:

$$|f(\boldsymbol{x})| \leq B \sum_{k=1}^{K} |F_k(\boldsymbol{x})| = B||\boldsymbol{F}(\boldsymbol{x})||_1.$$

Thus $G(\boldsymbol{x})$ is an envelope for $B\mathcal{F}$, and also $B||\boldsymbol{F}(\boldsymbol{x})||_1$ is an envelope for $\mathcal{F}_k$ for each $k \geq 1$. Now the class $\mathcal{G}_B$ is contained within the class:

$$\bar{\mathcal{G}}_B := \left\{ \sum_{k=1}^{K} t_k f_k(\boldsymbol{x}) : \sum_{k=1}^{K} t_k \leq B \text{ and } f_k \in \mathcal{F} \right\}.$$

Furthermore, $\bar{\mathcal{G}}_B$ is the symmetric convex hull of $B\mathcal{F}$; that is, if $\overline{\text{conv}}(B\mathcal{F})$ is the closed convex hull of $B\mathcal{F}$, then $\bar{\mathcal{G}}_B = \overline{\text{conv}}(B\mathcal{F}) \cup \overline{\text{conv}}(-B\mathcal{F}) \cup \{0\}$. Since $\mathcal{F}_1, \ldots, \mathcal{F}_K$ are VC-type classes by assumption, we have:

$$N(\epsilon ||F_k||_{Q,2}, \mathcal{F}_k, ||\cdot||_{Q,2}) \precsim \left(\frac{1}{\epsilon}\right)^{v_k},$$

for $v_k < \infty$ and $k = 1, \ldots, K$. Now note:

$$\begin{aligned}
N(\epsilon ||G||_{Q,2}, B\mathcal{F}, ||\cdot||_{Q,2}) &\leq N(\epsilon || B||\boldsymbol{F}||_1 ||_{Q,2}, B\mathcal{F}, ||\cdot||_{Q,2}) \\
&\leq N(\epsilon || ||\boldsymbol{F}||_1 ||_{Q,2}, \mathcal{F}, ||\cdot||_{Q,2}) \\
&\leq K \max_{1 \leq k \leq K} N(\epsilon || ||\boldsymbol{F}||_1 ||_{Q,2}, \mathcal{F}_k, ||\cdot||_{Q,2}) \\
&\leq K \max_{1 \leq k \leq K} N(\epsilon ||F_k||_{Q,2}, \mathcal{F}_k, ||\cdot||_{Q,2}) \precsim \left(\frac{1}{\epsilon}\right)^{v_{k^*}},
\end{aligned}$$

for all $0 < \epsilon \leq 1$, where k^* is the argmax of the left side of the last equality, and where the second

inequality uses Lemma 14.12 in Anthony and Bartlett (2009). Now set $v = 2v_{k^*}/(v_{k^*}+2)$, and note that $0 \leq v < 2$ when $0 \leq v_{k^*} < \infty$. By Theorem 2.6.9 in Van Der Vaart and Wellner (1996):

$$\log N(\epsilon||G||_{Q,2}, \overline{\text{conv}}(B\mathcal{F}), ||\cdot||_{Q,2}) \precsim \left(\frac{1}{\epsilon}\right)^v.$$

The result follows from the inequality:

$$N(\epsilon||G||_{Q,2}, \bar{\mathcal{G}}_B, ||\cdot||_{Q,2}) \leq 2N(\epsilon||G||_{Q,2}, \overline{\text{conv}}(B\mathcal{F}), ||\cdot||_{Q,2}) + 1,$$

and the fact that $N(\epsilon||G||_{Q,2}, \mathcal{G}_B, ||\cdot||_{Q,2}) \leq N(\epsilon||G||_{Q,2}, \bar{\mathcal{G}}_B, ||\cdot||_{Q,2})$.

■

Lemma S.5.3. *Suppose that $\mathcal{F} := \{f_\alpha(\cdot,\theta) : \mathcal{X} \to \mathbb{R} : \theta \in \Theta, \alpha \in \mathcal{A}\}$ is totally bounded parametric class of measurable real-valued functions on the metric space $(\mathcal{X}, d)$, where $(\mathcal{A}, d_a)$ and (Θ, d_θ) are also metric spaces. Furthermore let $\mathcal{G}$ be a class of real-valued functions with each element $g(\cdot,\theta) : \mathcal{X} \to \mathbb{R}$ defined by:*

$$g(\boldsymbol{x},\theta) := \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta),$$

for some $f \in \mathcal{F}$, where $C(\boldsymbol{x},\theta)$ is a nonempty multifunction for each $(\boldsymbol{x},\theta)$ pair. Then for any probability measure Q we have:

$$N(\epsilon, \mathcal{G}, ||\cdot||_{Q,2}) \leq N(\epsilon/2, \mathcal{F}, ||\cdot||_{Q,2}),$$

for any $0 < \epsilon < \infty$.

Proof of Lemma S.5.3. As a parametric class of functions (parameterized by (α,θ)), the $\epsilon/2-$cover of $\mathcal{F}$ can be characterized by a collection of points $\{(\alpha_i,\theta_i)\}_{i=1}^n$, where $n = N(\epsilon/2, \mathcal{F}, ||\cdot||_{Q,2})$. Denote such a collection by $\mathcal{N}(\mathcal{F})$. We will show that for any $g \in \mathcal{G}$ there exists a pair $(\alpha',\theta') \in \mathcal{N}(\mathcal{F})$ such that:

$$|g(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta')| \leq \epsilon.$$

Since every $g \in \mathcal{G}$ can be expressed as:

$$g(\boldsymbol{x},\theta) = \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta),$$

it suffices to show there exists a pair $(\alpha',\theta') \in \mathcal{N}(\mathcal{F})$ such that:

$$\left| \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta') \right| \leq \epsilon.$$

Now let α^* be any value satisfying:

$$\left| \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta) - f_{\alpha^*}(\boldsymbol{x},\theta) \right| \leq \epsilon/2.$$

That is, α^* is a $\epsilon/2$ solution to the minimization problem. Now choose the pair $(\alpha', \theta') \in \mathcal{N}(\mathcal{F})$ such that $|f_{\alpha^*}(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta')| \leq \epsilon/2$ (such a choice is always possible since $\mathcal{N}(\mathcal{F})$ is a $\epsilon/2-$cover of $\mathcal{F}$). Then we have:

$$\begin{aligned}
|g(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta')| &= \left| \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta') \right| \\
&\leq \left| \inf_{\alpha \in C(\boldsymbol{x},\theta)} f_\alpha(\boldsymbol{x},\theta) - f_{\alpha^*}(\boldsymbol{x},\theta) \right| + \left| f_{\alpha^*}(\boldsymbol{x},\theta) - f_{\alpha'}(\boldsymbol{x},\theta') \right| \\
&\leq \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

This completes the proof.

■

Lemma S.5.4. *Let $\mathcal{G}$ and $\mathcal{H}$ be two subsets of a normed vector space $(\mathfrak{X}, ||\cdot||)$, and let $\mathcal{F} := \mathcal{G} + \mathcal{H}$. Then:*

$$N(\epsilon, \mathcal{F}, ||\cdot||) \leq N(\epsilon/2, \mathcal{G}, ||\cdot||) N(\epsilon/2, \mathcal{H}, ||\cdot||),$$

for any $0 < \epsilon < \infty$.

Proof of Lemma S.5.4. Suppose that $N(\epsilon/2, \mathcal{G}, ||\cdot||) = n$ and $N(\epsilon/2, \mathcal{H}, ||\cdot||) = m$. It suffices to show $N(\epsilon, \mathcal{F}, ||\cdot||) \leq nm$. Let $\mathcal{N}(\mathcal{G})$ denote the centres of the balls that obtain the $n-$cover of $\mathcal{G}$ and let $\mathcal{N}(\mathcal{H})$ denote the centres of the balls that obtain the $m-$cover of $\mathcal{H}$. Enumerate the elements of $\mathcal{N}(\mathcal{G})$ as $g_1, \ldots, g_n$ and enumerate the elements of $\mathcal{N}(\mathcal{H})$ as $h_1, \ldots, h_m$. Now define the following collections:

$$G_j := \{g \in \mathcal{G} : ||g - g_j|| \leq \epsilon/2\}, \qquad H_k := \{h \in \mathcal{H} : ||h - h_k|| \leq \epsilon/2\},$$

for $j = 1, \ldots, n$ and $k = 1, \ldots, m$. For any $g_j \in \mathcal{N}(\mathcal{G})$ and $h_k \in \mathcal{N}(\mathcal{H})$ let $f_{jk} = g_j + h_k$, and define $F_{jk} := \{f : ||f - f_{jk}|| \leq \epsilon\}$. We argue that $\{F_{jk}\}$ is a $\epsilon-$cover of $\mathcal{F}$. Note that if we can establish this, the proof will be complete, since there are only nm sets F_{jk}. By construction each F_{jk} is a $||\cdot||-$ball of radius ϵ, so it only remains to check that $\{F_{jk}\}$ covers $\mathcal{F}$. To do so, fix any $f \in \mathcal{F}$. Then by definition $f = g + h$ for some $g \in \mathcal{G}$ and $h \in \mathcal{H}$. Since $\{G_j\}$ forms a $\epsilon/2-$cover of $\mathcal{G}$ and $\{H_k\}$ forms a $\epsilon/2-$cover of $\mathcal{H}$, we know there is some $g_j \in \mathcal{N}(\mathcal{G})$ and some $h_k \in \mathcal{N}(\mathcal{H})$ such that $||g - g_j|| \leq \epsilon/2$ and $||h - h_k|| \leq \epsilon/2$. But since $f_{jk} = g_j + h_k$ we have that:

$$||f - f_{jk}|| = ||(g + h) - (g_j + h_k)|| \leq ||g - g_j|| + ||h - h_k|| \leq \epsilon/2 + \epsilon/2 = \epsilon,$$

so that $f \in F_{jk}$, and so is an element of the cover $\{F_{jk}\}$. Since $f \in \mathcal{F}$ was arbitrary, we conclude that $\{F_{jk}\}$ covers $\mathcal{F}$. This completes the proof.

■

References

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

Anthony, M., and P. L. Bartlett (2009): *Neural network learning: Theoretical foundations*. cambridge university press.

Bauschke, H. H., P. L. Combettes, et al. (2011): *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408. Springer.

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

Bertsekas, D., A. Nedic, and A. Ozdaglar (2003): "Convex Analysis and Optimization," *Athena Scientific*.

Bertsekas, D. P. (1971): "Control of uncertain systems with a set-membership description of the uncertainty.," Ph.D. thesis, Massachusetts Institute of Technology.

Champion, T., and L. De Pascale (2011): "The Monge problem in R d," .

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Gaussian approximation of suprema of empirical processes," *The Annals of Statistics*, 42(4), 1564–1597.

Christensen, T., and B. Connault (2023): "Counterfactual sensitivity and robustness," *Econometrica*, 91(1), 263–298.

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

Dudley, R. M. (2014): *Uniform central limit theorems*. Cambridge university press.

Hörmander, L. (2007): *Notions of convexity*. Springer Science & Business Media.

Marcoux, M., T. M. Russell, and Y. Wan (2023): "A Simple Specification Test for Models with Many Conditional Moment Inequalities," *SSRN preprint ssrn.4345300*.

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

Rockafellar, R. T. (1970): *Convex analysis*, vol. 28. Princeton university press.

Ruder, S. (2016): "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*.

Van Der Vaart, A. W., and J. A. Wellner (1996): "Weak convergence," in *Weak convergence and empirical processes*, pp. 16–28. Springer.

Villani, C. (2003): *Topics in optimal transportation*, no. 58. American Mathematical Soc.