ONLINE SUPPLEMENTARY MATERIAL for "Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects"

Thomas Russell
Department of Economics, University of Toronto

October 10, 2019

1 Mathematical Preliminaries

This appendix reviews concepts from the theory of random sets that may assist the reader. Let $\mathcal{X}$ be a bounded subset of the $d-$dimensional euclidean space $\mathbb{R}^d$ and let $\mathcal{F}$ denote the set of closed sets on $\mathcal{X}$ and $\mathcal{K}$ denote the set of compact sets on $\mathcal{X}$.[1] Let $\mathscr{B}(\mathcal{F})$ be the $\sigma-$algebra generated by sets of the form $\{F : F \cap A \neq \emptyset\}$ for all compact $A \in \mathcal{K}$. Fix some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\mathbf{X} : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{F}, \mathscr{B}(\mathcal{F}))$.

Definition 1 (Random Closed Set (Molchanov (2005), pg. 1)). *The map $\mathbf{X} : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{F}, \mathscr{B}(\mathcal{F}))$ is called a random closed set if, for every compact set A in $\mathcal{X}$:*

$$\{\omega : \mathbf{X}(\omega) \cap A \neq \emptyset\} \in \mathscr{F}$$

Definition 2 (Capacity Functional (Molchanov (2005), pg. 4)). *A functional $T : \mathcal{K} \to [0, 1]$ given by*

$$T(A) = \mathbb{P}(\mathbf{X} \cap A \neq \emptyset), \qquad A \in \mathcal{K}$$

is called the capacity functional of the random set $\mathbf{X}$.

Since the random sets $\mathbf{X}$ and $\mathbf{X}'$ have realizations in the compact sets in $\mathbb{R}^d$, we have that $\mathbf{X}$ and $\mathbf{X}'$ are identically distributed (denoted $\mathbf{X} \overset{d}{\sim} \mathbf{X}'$) if and only if $\mathbb{P}(\mathbf{X} \cap A \neq \emptyset) = \mathbb{P}(\mathbf{X}' \cap A \neq \emptyset)$ for all $A \in \mathcal{K}$ (i.e. their capacity functionals agree for all compact sets). Note that, although $T(\emptyset) = 0$ and $T(\mathcal{U}) = 1$, unlike a typical probability measure the capacity functional T is generally non-additive. In particular, for two sets $A_1, A_2 \in 2^{\mathcal{U}}$ such that $A_1 \cap A_2 = \emptyset$ we may have:

$$\{\mathbf{X} \cap A_1 \neq \emptyset\} \cap \{\mathbf{X} \cap A_2 \neq \emptyset\} \neq \emptyset,$$

[1] Note that since we consider a bounded subset $\mathcal{X} \subset \mathbb{R}^d$, all closed sets on $\mathcal{X}$ are compact.

which implies

$$T(A_1 \cup A_2) < T(A_1) + T(A_2).$$

An important concept in random set theory is the idea of a *selection* of a random set, which can be intuitively understood as a random variable with realizations within the random set:

Definition 3 (Selection, Molchanov (2005) pg. 26)**.** *A random variable $X : (\Omega, \mathscr{F}) \to (\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is called a (measurable) selection of the random set $\mathbf{X}$ if $X \in \mathbf{X}$ $\mathbb{P}$-a.s. The family of all selections of $\mathbf{X}$ is denoted $sel(\mathbf{X})$.*

In the context of this paper, we are particularly interested in the measurable selections U from the random set $G^{-1}(W)$. With this terminology, the following Theorem leads directly to the key identification results in this paper:

Theorem (Artstein's Theorem)**.** *Let X be a random variable with distribution μ and let $\mathbf{X}$ be a random set with distribution ν. Then there exists a random variable X' and a random set $\mathbf{X}'$ with $X' \overset{d}{\sim} X$ and $\mathbf{X}' \overset{d}{\sim} \mathbf{X}$ such that $X' \in sel(\mathbf{X}')$ if and only if:*

$$\mu(X \in A) \leq \nu(\mathbf{X} \cap A \neq \emptyset) \qquad \forall A \in \mathcal{K}(\mathcal{R}^d) \tag{1}$$

2 Core Determining Classes for Treatment Effects

The Exact Core Determining Class

Luo and Wang (2016) define the *exact core determining class* as the smallest core determining class. This fact motivates the following definition from Luo and Wang (2016):

Definition 4 (Luo and Wang (2016))**.** *The exact core determining class $\mathcal{S}^*$ is the collection of all subsets $A \in 2^{\mathcal{U}}$ and $A \neq \mathcal{U}$ such that*

$$Q^*(A) > P(G^{-1}(Y, D) \cap A \neq \emptyset)$$

where

$$Q^*(A) \equiv \max\{Q(A) | Q(A') \leq P(G^{-1}(Y, D) \cap A' \neq \emptyset) \ \forall A' \in 2^{\mathcal{U}}, \ A' \neq A; \ Q(\mathcal{U}) = 1\}.$$

As the results in this appendix show, thinking about the exact core determining class in terms of non-redundant linear inequality constraints is convenient. To facilitate comparison with results that appear later, we restate the technical result of Luo and Wang (2016) here. First, a definition of important set collections that can be used to characterize the exact core determining class.

Definition 5 (Luo and Wang (2016))**.** *Let $\mathcal{S}_u$, $\mathcal{S}_w$ and $\mathcal{S}_w^{-1}$ be the collections of sets with the following properties:*

(a) $\mathcal{S}_u$ is the collection of all non-empty subsets $A \in 2^{\mathcal{U}}$, $A \neq \mathcal{U}$, such that

(i) *A is self-connected.*[2]

(ii) *There exists no $u \in \mathcal{U}$ such that $u \notin A$ and $G(u) \subset G(A)$.*

(b) *$\mathcal{S}_w$ is the collection of all non-empty subsets $B \in 2^{\mathcal{W}}$, $B \neq \mathcal{W}$, such that*

(i) *B is self-connected.*

(ii) *There exists no $w \in \mathcal{W}$ such that $w \notin B$ and $G^{-1}(w) \subset G^{-1}(B)$.*

(c) *$\mathcal{S}_w^{-1}$ is the collection of $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$ such that there exists $B \subset \mathcal{S}_w$ such that $A = G^{-1}(B)^c$.*

Note that condition (i) in the definition of both $\mathcal{S}_u$ and $\mathcal{S}_w$ corresponds to the redundancy condition suggested by Chesher and Rosen (2017). Condition (ii) in the definition of both $\mathcal{S}_u$ and $\mathcal{S}_w$ is novel to the paper by Luo and Wang (2016). Intuitively, $\mathcal{S}_u$ and $\mathcal{S}_w$ represent the collection of non-redundant sets when Artstein's inequalities are defined on the unobservables and observables, respectively. Furthermore, the collection $\mathcal{S}_w^{-1}$ is the "reflection" in the space of unobservables of the non-redundant sets in the space of observables. The main result in Luo and Wang (2016) follows.

Theorem (Luo and Wang (2016))**.** *Assume that $\mathcal{G}$ is self-connected. If the measure $\mathcal{P}$ on $\mathcal{W}$ is non-degenerate, i.e. $\mathcal{P}(w)$ is non-zero for all $w \in \mathcal{W}$, then the exact core determining class is given by:*

$$\mathcal{S}^* = \mathcal{S}_u \cap \mathcal{S}_w^{-1}$$

Using this result, Luo and Wang (2016) provide an algorithm to compute the exact core determining class for a general econometric model and provide some Monte Carlo evidence showing that the exact core determining class is able to reduce the number of inequalities significantly.[3] Intuitively, to find the core determining class we must:

(i) Decide which sets $A \in 2^{\mathcal{U}}$ satisfy the conditions necessary to belong to $\mathcal{S}_u$.

(ii) Decide which sets $A' \in 2^{\mathcal{W}}$ satisfy the conditions necessary to belong to $\mathcal{S}_w$.

(iii) Decide which sets $A \in 2^{\mathcal{U}}$ satisfy the conditions necessary to belong to $\mathcal{S}_w^{-1}$.

(iv) Intersect the sets $\mathcal{S}_u$ and $\mathcal{S}_w^{-1}$.

Since the number of sets in $2^{\mathcal{U}}$ and $2^{\mathcal{W}}$ can be prohibitively large, even an efficient algorithm can take an unreasonable amount of time to characterize the exact core determining class.

Note that the POM provides a very specific structure to the correspondence G. The structure of the correspondence G in the POM is best illustrated when looking at the bipartite graph $\mathcal{G} = (\mathcal{W}, \mathcal{U}, G)$. Some appealing properties of the general bipartite graph $\mathcal{G}$ defined by the POM include:

[2] A set A is self-connected if for every $A_1, A_2 \subset A$ such that $A_1, A_2 \neq \emptyset$ and $A_1 \cup A_2 = A$ we have $G(A_1) \cap G(A_2) \neq \emptyset$.

[3] Luo and Wang (2017) mention that example 3 in Luo and Wang (2016) is able to eliminate 98.56% of the inequalities in a 15×25 bipartite graph.

(i) Part $\mathcal{U}$ of the graph $\mathcal{G}$ has exactly $|\mathcal{Y}|^{|\mathcal{D}|}$ nodes with degree $|\mathcal{D}|$.

(ii) Part $\mathcal{W}$ of the graph $\mathcal{G}$ has exactly $|\mathcal{Y}||\mathcal{D}|$ nodes with degree $|\mathcal{Y}|^{|\mathcal{D}|-1}$.

(iii) For $u_1 \neq u_2$, we have $G(u_1) \neq G(u_2)$. Similarly, for $w_1 \neq w_2$, we have $G^{-1}(w_1) \neq G^{-1}(w_2)$.

(iv) $\mathcal{G}$ is connected.

Using the properties of the graph $\mathcal{G}$, it is possible to characterize the properties of the sets in the exact core determining class for the POM. Results on the precise nature of sets in the exact core determining class in the POM are given in Lemmas 1, 2 and 3 below.

Lemma 1. *For the POM, $A \in \mathcal{S}_u$ and $|A| \geq 2$ if and only if all singletons that comprise A have exactly $|\mathcal{D}| - 1$ elements in common.*

Lemma 2. *For the POM we have*

(a) $\mathcal{G}$ can be partitioned into $|\mathcal{D}|$ disjoint subgraphs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{|\mathcal{D}|}$ with $\mathcal{G}_k = (\mathcal{W}_k, \mathcal{U}, G)$, where

(i) $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ for all $i \neq j$.

(ii) $G^{-1}(w) \cap G^{-1}(w') \neq \emptyset$ for any pair (w, w') with $w \in \mathcal{W}_i$, $w' \in \mathcal{W}_j$, $i \neq j$.

(iii) $G^{-1}(w) \cap G^{-1}(w') = \emptyset$ for any $w, w' \in \mathcal{W}_k$.

(iv) $G^{-1}(\mathcal{W}_k) = \mathcal{U}$ for every k.

(b) $B \in \mathcal{S}_w$ if and only if:

(i) $B \not\subseteq \mathcal{W}_k$ for any k if $|B| \geq 2$.

(ii) $\mathcal{W}_k \not\subseteq B$ for any k.

Lemma 3. *If $|\mathcal{D}| = 2$ and $|\mathcal{D}| < |\mathcal{Y}|$, then $\mathcal{S}_w^{-1}$ contains all sets $A \subset \mathcal{S}_u$ with $|A| \leq |\mathcal{Y}| - 1$. Otherwise, $\mathcal{S}_u \subset \mathcal{S}_w^{-1}$.*

To summarize, Lemmas 1 and 3 provide a complete characterization of the type of sets in the exact core determining class, and Lemma 2 provides information on the structure of the POM bipartite graph. Further intuition on the interpretation of sets selected the exact core determining class is provided in the main paper. These Lemmas can then be used to prove the following result, which was presented in the main text.

Theorem 1. *Suppose that the distribution P is non-degenerate:*

1. *In the POM there are exactly:*

$$\begin{cases} |\mathcal{Y}|^{|\mathcal{D}|} & \text{if } r = 1 \\ |\mathcal{Y}|^{|\mathcal{D}|-1}|\mathcal{D}| \cdot \binom{|\mathcal{Y}|}{r} & \text{if } r \geq 2 \end{cases}$$

r-element sets in the collection $\mathcal{S}_u$.

2. In the POM there are exactly:

$$\sum_{\ell=2}^{|\mathcal{D}|} \binom{|\mathcal{D}|}{\ell} \left(\sum_{v \in A(r,|\mathcal{Y}|,\ell)} \prod_{i=1}^{\ell} \binom{|\mathcal{Y}|}{v_i} \right)$$

r-element sets in the collection $\mathcal{S}_w$, where

$$A(r, |\mathcal{Y}|, \ell) = \left\{ (v_1, v_2, \ldots, v_\ell) \in \mathbb{N}^\ell : \sum_i v_i = r, \quad 1 \leq v_i \leq |\mathcal{Y}| - 1 \, \forall i \right\}$$

3. In the POM there are

$$\begin{cases} |\mathcal{Y}|^{|\mathcal{D}|} + \sum_{r=2}^{|\mathcal{Y}|} |\mathcal{Y}|^{|\mathcal{D}|-1} |\mathcal{D}| \binom{|\mathcal{Y}|}{r} - |\mathcal{Y}||\mathcal{D}|, & \text{if } |\mathcal{D}| = 2 \text{ and } |\mathcal{Y}| > |\mathcal{D}| \\ |\mathcal{Y}|^{|\mathcal{D}|} + \sum_{r=2}^{|\mathcal{Y}|} |\mathcal{Y}|^{|\mathcal{D}|-1} |\mathcal{D}| \binom{|\mathcal{Y}|}{r}, & \text{otherwise} \end{cases}$$

sets in the exact core determining class.

3 Conditional Probability/Linear Programming

This Appendix gives an example of how to implement the optimization problems suggested in Theorem 1. Suppose for simplicity that we are in the binary outcome, binary treatment case. Let $q_{ij} = \mathbb{P}(Y_0 = i, Y_1 = j)$, and suppose we wish to bound the parameter

$$\mathbb{P}(Y_1 = 1 | Y_0 = 0) = \frac{q_{01}}{q_{00} + q_{01}}$$

It is possible to show that we can bound this parameter using a linear program. First note that we can write the dual problem to Artstein's inequalities (discussed in Section 3) as:

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}}_{\mathbf{A}_\pi} \underbrace{\begin{bmatrix} \pi_{00,0} \\ \pi_{01,0} \\ \pi_{10,0} \\ \pi_{11,0} \\ \pi_{00,1} \\ \pi_{01,1} \\ \pi_{10,1} \\ \pi_{11,1} \end{bmatrix}}_{\pi} = \underbrace{\begin{bmatrix} p_{00} \\ p_{01} \\ p_{10} \\ p_{11} \end{bmatrix}}_{\mathbf{p}}$$

which trivially impose only linear constraints. Also recall that we can write:

$$\begin{aligned} q_{00} &= \pi_{00,0} + \pi_{00,1} \\ q_{01} &= \pi_{01,0} + \pi_{01,1} \\ q_{10} &= \pi_{10,0} + \pi_{10,1} \end{aligned}$$

$$q_{11} = \pi_{11,0} + \pi_{11,1}$$

Then the optimization problem is:

$$\max_{\pi} \frac{\pi_{01,0} + \pi_{01,1}}{\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}} \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \pi = \mathbf{p} \\ \mathbf{0} \preccurlyeq \pi \preccurlyeq \mathbf{1} \end{cases} \tag{2}$$

To write this as a linear programming problem, define

$$r = \frac{1}{\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}}, \qquad \tilde{\pi} = \begin{bmatrix} \pi_{00,0}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{01,0}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{10,0}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{11,0}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{00,1}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{01,1}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{10,1}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \\ \pi_{11,1}/(\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}) \end{bmatrix}$$

$$c = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad d_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad d_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Then the problem above can be re-written

$$\max_{\tilde{\pi}, r} c' \cdot \tilde{\pi} \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \tilde{\pi} - \mathbf{p} \cdot r = 0 \\ d_1 \cdot \tilde{\pi} = 1 \\ d_2 \cdot \tilde{\pi} - r = 0 \\ \mathbf{0} \preccurlyeq \tilde{\pi} \preccurlyeq \mathbf{1} \\ r \geq 1 \end{cases} \tag{3}$$

This can be seen by replacing the objective function in (2) with the equivalent objective function in (3), by multiplying both sides of the constraint $\mathbf{A}_\pi \cdot \pi = \mathbf{p}$ in (2) by the variable r and rearranging, and by imposing constraints ensuring that the conditional probability measure is a proper probability measure, namely:

$$
\begin{aligned}
d_1 \cdot \tilde{\pi} = 1 \qquad &\Longrightarrow \qquad \sum_j \mathbb{P}(Y_1 = y_j | Y_0 = 0) = 1 \\
d_2 \cdot \tilde{\pi} - r = 0 \qquad &\Longrightarrow \qquad \sum_i \sum_j \mathbb{P}(Y_0 = y_i, Y_1 = y_j) = 1 \\
\mathbf{0} \preccurlyeq \tilde{\pi} \preccurlyeq \mathbf{1} \text{ and } r \geq 0 \qquad &\Longrightarrow \qquad 0 \leq \mathbb{P}(Y_0 = y_i, Y_1 = y_j) \leq 1 \qquad \forall i, j
\end{aligned}
$$

Alternatively, we could write the same problem more compactly as

$$
\max_{\tilde{\mathbf{q}}_r} c_r' \cdot \tilde{\mathbf{q}}_r \qquad s.t. \qquad \begin{cases} \mathbf{A}_r \cdot \tilde{\mathbf{q}}_r = \mathbf{a}_r \\ b_l \preccurlyeq \tilde{\mathbf{q}}_r \preccurlyeq b_u \end{cases} \tag{4}
$$

where $\tilde{\mathbf{q}}_r' = (\tilde{\pi}', r)'$ and where

$$
\mathbf{A}_r = \begin{bmatrix} \mathbf{A}_\pi & -\mathbf{p} \\ d_1' & 0 \\ d_2' & -1 \end{bmatrix}, \qquad \mathbf{a}_r = \begin{bmatrix} \mathbf{0} \\ 1 \\ 0 \end{bmatrix},
$$

$$
c_r = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad b_l = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad b_u = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \infty \end{bmatrix}.
$$

The problem (4) is now in a form amenable for implementation in common linear programming software; for example, Matlab and Gurobi. It is also easily generalized to cases beyond binary treatment and binary outcome.

3.1 Introducing Additional Constraints

Imposing additional assumptions on the unobserved probability measure Q in an analytic framework requires a new proposed identified set and corresponding proof of sharpness. In contrast, additional assumptions can be imposed easily on Q in the computational frame-

work. In addition, in many cases additional assumptions can be included as linear constraints in Q, which are convenient from a computational point of view.

Additional constraints are often useful when the identified set for a parameter of interest is wide, as introducing constraints on Q can result in a more informative identified set. These additional constraints allow a researcher to trade-off the length of the bounds with the credibility of the maintained assumptions. Perhaps the most well-known assumptions used in the partial identification of treatment effects are the monotone treatment response (MTR) assumption and the monotone instrumental variables assumption (MIV), which are outlined in Manski and Pepper (2000) and discussed in Manski (2003).

Definition 6 (MTR, Manski and Pepper (2000))**.** *Let $\mathcal{Y}_d$ be an ordered set. Then the MTR assumption is satisfied if $d' \geq d \implies \mathbb{P}(Y_{d'} \geq Y_d) = 1$.*

I.e. the MTR assumption implies that the potential outcomes are monotone in the treatment, and can be useful when a researcher has some strong *a priori* evidence that a particular treatment is effective at increasing (decreasing) an outcome variable Y for all individuals. It is also possible to order potential outcomes with respect to a variable other than treatment status, which motivates the MIV assumption:

Definition 7 (MIV, Manski and Pepper (2000))**.** *Suppose that $\mathcal{Z}$ is an ordered set. The covariate Z is a monotone instrumental variable if for each treatment $d \in \mathcal{Y}_d$, we have that $z' \geq z \implies \mathbb{E}[Y_d|Z = z'] \geq \mathbb{E}[Y_d|Z = z]$.*

Note that the MTR and MIV assumptions can be written as constraints on the unobserved probability measure Q. Indeed, it has been shown by Demuynck (2015), Lafférs (2013, 2015) and Torgovitsky (2016) that these assumptions, and versions thereof, can be written as linear constraints on Q (which makes them especially amenable to inclusion in linear programs). Since the set $\mathcal{Q}^\dagger$ is still convex and closed under these constraints, estimation using Artstein's inequalities is consistent by Theorem 2. The MTR and MIV assumptions presented are examples of additional assumptions that can be imposed to obtain a more

informative analysis, although there are many other assumptions that might also be imposed without affecting any of the previous results.

4 Consistency and Inference

In this section we show the conditions under which the optimization-based bounding procedure is consistent, and we repeat some discussion given in the main paper. Consistency is presented without an instrument for simplicity, but the result also holds when a instrument with finite support is available. Finally, the proof of consistency is given for the case when $\mathcal{Q}$ is defined by Artstein's inequalities rather than the dual approach, although it is applicable to both approaches since both approaches give numerically identical characterizations of $\mathcal{Q}$.

Consider the usual empirical measure:

$$\mathbb{P}_n(A) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{(Y_i, D_i) \in A\},$$

and define the sets

$$\mathcal{Q}(\mathbb{P}_n) \equiv \{Q \in \mathcal{Q}^\dagger : Q(A) \leq \mathbb{P}_n(G^{-1}(Y, D) \cap A \neq \emptyset) \text{ for all } A \in 2^{\mathcal{U}}\},$$

or equivalently:

$$\mathcal{Q}(\mathbb{P}_n) \equiv \{Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}(\mathbb{P}_n, Q)\}.$$

Consistency in the estimation of sets is usually defined in terms of the Hausdorff distance d_H, which furnishes a metric on the space of non-empty compact subsets of $\mathbb{R}^d$.[4] Here we are interested in establishing consistency with respect to the Hausdorff metric of the set

$$\Theta^f(\mathbb{P}_n) = [f^\ell(\mathbb{P}_n), f^u(\mathbb{P}_n)] \quad \text{with} \quad f^\ell(\mathbb{P}_n) = \sup_{Q \in \mathcal{Q}(\mathbb{P}_n)} f(Q), \quad f^u(\mathbb{P}_n) = \inf_{Q \in \mathcal{Q}(\mathbb{P}_n)} f(Q) \tag{5}$$

for the set

$$\Theta^f(P) = [f^\ell(P), f^u(P)] \quad \text{with} \quad f^\ell(P) = \sup_{Q \in \mathcal{Q}(P)} f(Q), \quad f^u(P) = \inf_{Q \in \mathcal{Q}(P)} f(Q) \tag{6}$$

[4] The Hausdorff distance for any two sets A and B as:

$$d_H(A, B) = \max\left\{\sup_{a \in A}\inf_{b \in B}||a - b||, \sup_{b \in B}\inf_{a \in A}||a - b||\right\}.$$

Consistency is given in the following Theorem, which is also presented in the main text:

Theorem. *Fix any continuous functional $f : \mathcal{Q} \to \mathbb{R}$. Suppose that (a) $\mathcal{Q}^\dagger$ is restricted only through linear (in)equality constraints; (b) the Jacobian of the linear equality constraints defining $\mathcal{Q}$ (if any) has full row rank; (c) $\{W_i\}_{i=1}^n$ is i.i.d. from some probability measure P with finite support; and (d) int $(\mathcal{Q}(P)) \neq \emptyset$. Then $\Theta^f(\mathbb{P}_n) \xrightarrow{p} \Theta^f(P)$ in the Hausdorff metric.*

Since f is a continuous functional, consistency follows if we can show that $\mathcal{Q}(\mathbb{P}_n) \xrightarrow{p} \mathcal{Q}(P)$ in the Hausdorff metric (see the proof for a detailed discussion). To begin the proof, we first show $\mathcal{Q}(\mathbb{P}_n)$ can be written as the set minimizer of an appropriately defined criterion function, as well-known consistency results exist for problems of this kind (see in particular Chernozhukov et al. (2007), Yildiz (2012), Menzel (2014) and Shi and Shum (2015)). The proof then follows by verifying that the problem fits into the framework of Shi and Shum (2015), and by verifying the conditions required for consistency presented in their paper.

Condition (a) in the Theorem is made primarily for simplicity, but also since it covers all the cases discussed in this paper. It is possible to relax condition (a), although it will then generally be harder to verify condition (b) if the linear equality constraints become non-linear equality constraints, since the gradients of these equality constraints would then depend on the parameter Q. Condition (b) in the Theorem is required to apply the consistency result of Shi and Shum (2015), and condition (c) is standard.

Condition (d) is worth some discussion. Note that Theorem 2 shows that estimation of bounds on any continuous functional of the joint distribution can be completed using Artstein's inequalities without the need for a tuning parameter. However, this is done at the cost of ruling out point identification through assumption (d). While point identification is a knife-edge case under all assumptions considered in this paper, some researchers may feel assumption (d) is too restrictive. If this is the case, researchers can add a slackness term that drifts towards zero —say c_n— to each of the inequalities defining the set $\mathcal{Q}$, and Theorem 2 can then be applied with assumption (d) replaced with the assumption that $\mathcal{Q} \neq \emptyset$. A general rule for selecting the slackness is that it should dominate relative to sampling error;

thus, a possible choice for the slackness is given by $c_n = \sqrt{\log(n)/n}$. Introducing such a slackness term will cause any estimated identified sets to have slightly larger length, although any difference will be negligible for large n.

5 Application Robustness Exercise

Figure 1 shows plots of $\mathbb{P}(Y_1 > y_q | Y_0 \leq y_{0.5})$ and $\mathbb{P}(Y_1 > y_{0.5} | Y_0 \leq y_q)$ against y_q, where y_q is the q^{th} quantile of the observed grade 3 ranks. The figures emphasize that, for the most part, the bounds on the conditional probability for the Tennessee STAR application are wide and uninformative. In contrast, Figure 2 shows informative plots of the joint distribution $\mathbb{P}(Y_1 > y_q, Y_0 \leq y_{0.5})$ and $\mathbb{P}(Y_1 > y_{0.5}, Y_0 \leq y_q)$ against y_q.

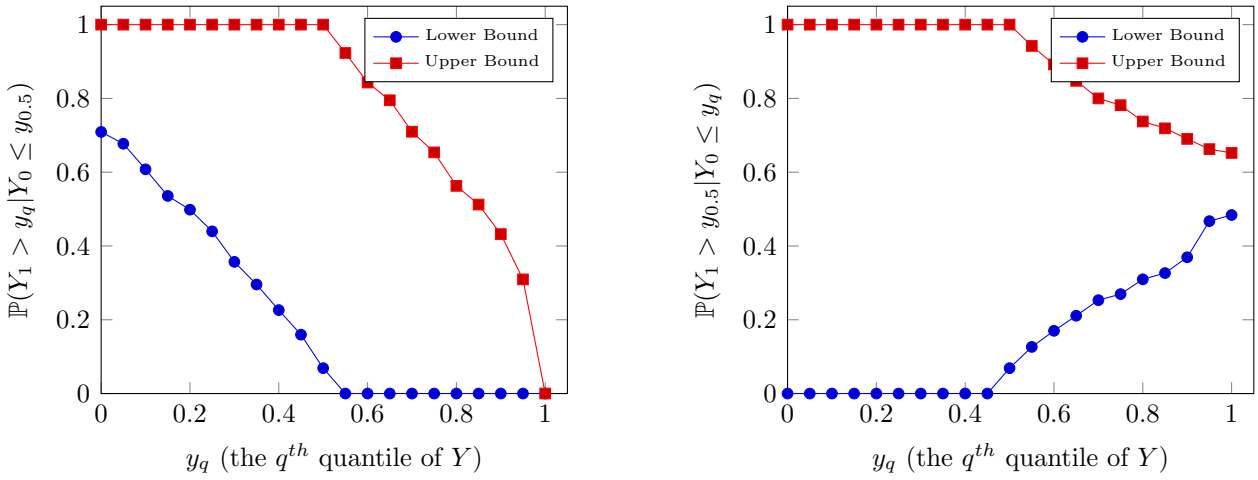

Figure 1: Bounds on the conditional probability (Grade 3, Bins=35, MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$).

Table 1 shows bounds for the parameters of interest in the Tennessee STAR experiment when the MTR condition is relaxed from $\mathbb{P}(Y_1 > Y_0) \geq 0.95$ to the MTR condition $\mathbb{P}(Y_1 > Y_0) \geq 0.5$. As discussed in the main text, the bounds on some of the parameters —such as the bounds on $\mathbb{P}(Y_1 > Median | Y_0 \leq Median)$, $\mathbb{P}(Y_0 \leq Median)$, $\sqrt{Var(Y_0)}$)— are almost completely unaffected by the relaxing of the assumption. However, bounds on

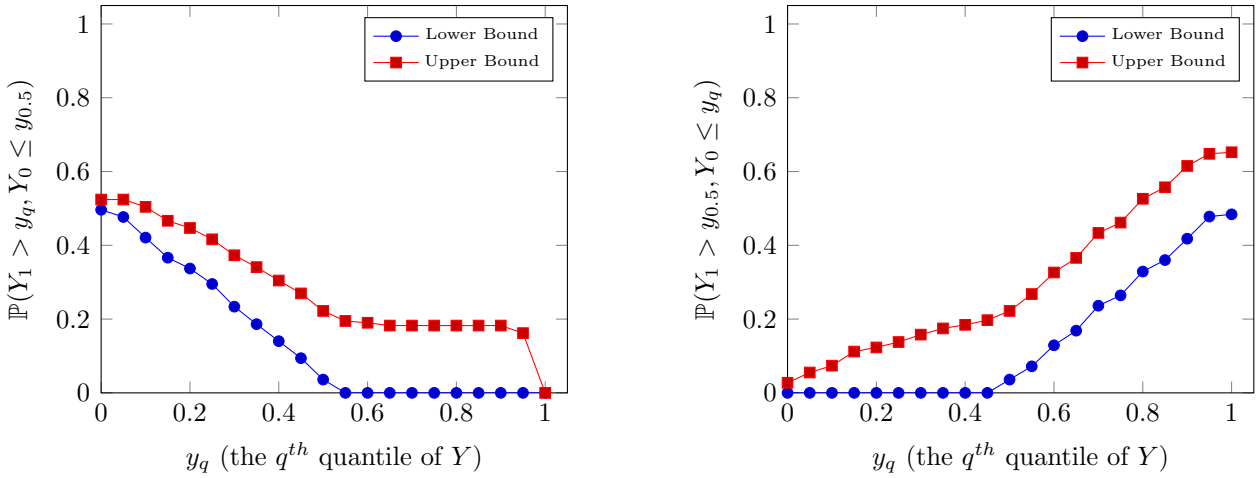

Figure 2: Bounds on the joint probability (Grade 3, Bins=35, MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$).

other parameters —especially $\mathbb{E}[Y_1 - Y_0]$ and $Corr(Y_0, Y_1)$— become uninformative when the assumption is relaxed. However, the reader is encouraged to keep in mind that under either condition ($\mathbb{P}(Y_1 > Y_0) \geq 0.95$ or $\mathbb{P}(Y_1 > Y_0) \geq 0.5$) the bounds are *sharp* in the sense that they exhaust all the information provided by the data under the maintained assumptions. Thus, whether the bounds are informative —and under which assumptions the bounds are informative— depends always on the empirical context, and not on the method proposed in this paper (which always delivers sharp bounds).

Table 1: Bounds on School Achievement from the Tennessee STAR Experiment Assuming $\mathbb{P}(Y_1 > Y_0) \geq 0.5$

		Y = Grade 3 percentile rank D = Small class K-3		Y = Grade 8 percentile rank D = Small class K-3	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
$\mathbb{P}(Y_0 \leq Median(Y), Y_1 > Median(Y))$:†	Bins=25	0.08	0.53	0.04	0.52
	Bins=30	0.08	0.53	0.04	0.52
	Bins=35	0.09	0.53	0.04	0.52
$\mathbb{E}[Y_1 - Y_0]$:	Bins=25	-6.26	19.27	-8.51	17.03
	Bins=30	-6.58	19.13	-9.39	16.32
	Bins=35	-7.52	18.32	-9.57	16.27
$\mathbb{P}(Y_1 > Y_0)$:*	Bins=25	0.11	0.97	0.05	0.97
	Bins=30	0.11	0.98	0.05	0.98
	Bins=35	0.11	0.98	0.05	0.98
$\mathbb{P}(Y_1 > Median(Y) \vert Y_0 \leq Median(Y))$:†	Bins=25	0.14	0.97	0.07	1.00
	Bins=30	0.14	0.96	0.07	1.00
	Bins=35	0.17	1.00	0.07	1.00
$\mathbb{P}(Y_0 \leq Median(Y))$:†	Bins=25	0.55	0.55	0.52	0.52
	Bins=30	0.55	0.55	0.52	0.52
	Bins=35	0.53	0.53	0.52	0.52
$\mathbb{P}(Y_1 > Median(Y))$:†	Bins=25	0.40	0.66	0.39	0.66
	Bins=30	0.39	0.66	0.39	0.65
	Bins=35	0.42	0.69	0.39	0.65
$Corr(Y_0, Y_1)$:	Bins=25	-0.50	0.50	-0.50	0.50
	Bins=30	-0.50	0.50	-0.50	0.50
	Bins=35	-0.50	0.50	-0.50	0.50
$\sqrt{Var(Y_1 - Y_0)}$:	Bins=25	2.37	43.90	0.84	42.75
	Bins=30	2.43	44.57	0.55	43.02
	Bins=35	2.07	43.89	0.89	45.26

†: Recall that Median(Y) is the median of the observed outcome, but not necessarily the median of Y_0 or Y_1.
∗: The parameter $\mathbb{P}(Y_1 > Y_0)$ is the only parameter estimated without the MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$.

6 Proofs

Proof of Theorem 1. Recall our probability space is $(\Omega, \mathscr{F}, \mathbb{P})$. Note since $\mathcal{U}$ is finite, then so is $G^{-1}(Y, D)$ since G^{-1} maps within $\mathcal{U}$. Since $\{(y, d) : G^{-1}(y, d) \cap A \neq \emptyset\} \in 2^{\mathcal{W}}$ for all $A \in 2^{\mathcal{U}}$, and since Y and D are random variables (measurable by assumption) we have that $\{\omega : G^{-1}(Y(\omega), D(\omega)) \cap A \neq \emptyset\} \in \mathscr{F}$ for all $A \in 2^{\mathcal{U}}$, and thus $G^{-1}(Y, D)$ is a random closed set. By Artstein's Theorem we have that for the random set $G^{-1}(Y, D)$ and for the element $U \in \mathcal{U}$, there exists a random set $[G']^{-1}(Y, D)$ and a random variable $U' \in \mathcal{U}$ such that

$[G']^{-1}(Y,D) \overset{d}{\sim} G^{-1}(Y,D)$ and $U' \overset{d}{\sim} U$ and $U' \in [G]^{-1}(Y,D)$ a.s. if and only if

$$\mathbb{P}(U \in A) \leq \mathbb{P}(G^{-1}(Y,D) \cap A \neq \emptyset) \qquad \forall A \in 2^{\mathcal{U}}$$

Thus, the collection $\mathcal{Q}$ provides a sharp characterization of the set of all joint distributions Q of $U \in \mathcal{U}$ consistent with the observed distribution P. If $\mathcal{Q}^\dagger$ is convex then $\mathcal{Q}$ is also convex, as it restricts $\mathcal{Q}^\dagger$ only via the linear inequality constraints implied by Artstein's Theorem. The result than follows from the proof of proposition 1 in Torgovitsky (2016). In particular, because $\mathcal{U}$ is finite with dimension $d_{\mathcal{U}}$, we have that $\mathcal{Q} \subset \mathbb{R}^{d_{\mathcal{U}}}$ is compact. Finally, the image of a continuous functional over a non-empty compact and convex set $\mathcal{Q} \subset \mathbb{R}^{d_{\mathcal{U}}}$ is a non-empty interval with the end points defined as in equation (9). ■

Proof of Lemma 1. For notational simplicity, let $M \equiv |\mathcal{Y}|$ and $K \equiv |\mathcal{D}|$.

First consider the reverse; i.e. suppose that A is a union of r singletons that have exactly $K-1$ elements in common. Note that for every pair of singletons $u, u' \in A$, we have $G(u) \cap G(u') \neq \emptyset$ and $G(u) \neq G(u')$. Thus, for any partition A_1, A_2 of A we always have $G(A_1) \cap G(A_2) \neq \emptyset$. Next, suppose by way of contradiction that there exists a $u \notin A$ such that $G(u) \subset G(A)$. Since $G(u) \subset G(A)$, it must be that u must have the same $K-1$ elements in common with all members of A (otherwise it cannot map within $G(A)$). However, since $u \notin A$ it must be that u has one element uncommon to all members of A. But then $G(u) \not\subset G(A)$, which gives the desired contradiction and completes the proof of the reverse direction.

Now consider the forward direction; i.e. suppose that $A \in \mathcal{S}_u$ and $|A| = r \geq 2$, and proceed by inducting on r. First consider the case when $r = 2$. For any $A \in \mathcal{S}_u$ with $|A| = 2$, take the singletons u_1, u_2 that comprise A (i.e. the singletons such that $u_1 \cup u_2 = A$). If u_1 and u_2 share more than $K-1$ elements then they are the same vector. It is also clear that u_1 and u_2 must share at least one element, otherwise condition (a)(i) in Definition 5 is not

satisfied. Thus, suppose u_1 and u_2 share $1 \leq k < K-1$ elements. Without loss of generality, suppose that they share the first k elements, so that we can write the vectors u_1 and u_2 as:

$$u_1 = (y_1, y_2, \ldots, y_k, y_{1(k+1)}, y_{1(k+2)}, \ldots, y_{1K})$$
$$u_2 = (y_1, y_2, \ldots, y_k, y_{2(k+1)}, y_{2(k+2)}, \ldots, y_{2K})$$

Now consider the vector u_3 given by:

$$u_3 = (y_1, y_2, \ldots, y_k, y_{1(k+1)}, y_{1(k+2)}, \ldots, y_{1(K-1)}, y_{2K})$$

I.e. u_3 is the vector that shares the same first k elements with both u_1 and u_2, shares the next $(K-1)-(k+1)$ elements with vector u_1, and shares the last element with vector u_2. Clearly this vector u_3 exists, $u_3 \notin A$ and $G(u) \subset G(u_1 \cup u_2)$, contradicting the fact that $A = u_1 \cup u_2$ is in $\mathcal{S}_u$. Thus we conclude that the claim holds for the base case of $r = 2$.

Now suppose the claim holds for $r = \ell$. Then we know that any $A \in \mathcal{S}_u$ such that $|A| = \ell$ must be comprised of singletons $u_1, u_2, \ldots, u_\ell$ that share $K-1$ elements. Without loss of generality suppose that these are the first $K-1$ elements so that we can write:

$$u_1 = (y_1, y_2, \ldots, y_{K-1}, y_{1K})$$
$$u_2 = (y_1, y_2, \ldots, y_{K-1}, y_{2K})$$
$$\vdots$$
$$u_\ell = (y_1, y_2, \ldots, y_{K-1}, y_{\ell K})$$

where $y_{iK} \neq y_{jK}$ for any $i \neq j$. Now consider a set $A' \in \mathcal{S}_u$ with $|A'| = \ell + 1$. Note that any such set can be constructed by adding a singleton u to a set $A \in \mathcal{S}_u$ where $|A| = \ell$, so that $A' = A \cup u$ for some $u \in \mathcal{U}$. Thus, suppose by way of contradiction that there exists a $u_{\ell+1} \in \mathcal{U}$ such that for some $A \in \mathcal{S}_u$ we have $A' = A \cup u_{\ell+1} \in \mathcal{S}_u$, but that $u_{\ell+1}$ does not have $K-1$ elements in common with every vector in A. Clearly $u_{\ell+1}$ cannot have more than $K-1$ elements in common with any vector in A, since then it is the same as one vector in A. Thus it must be that $u_{\ell+1}$ has less than $K-1$ elements in common with at least one vector

in A. Also note that clearly $u_{\ell+1}$ has at least one element in common with one vector $u_i \in A$ (otherwise A does not satisfy condition 1 in Definition 5). Suppose without loss of generality that this vector is $u_i = u_1$; this simplification is only to reduce the level of abstraction. Now consider two cases:

1. $u_{\ell+1}$ and u_1 share the element y_{1K}: the fact they share y_{1K} implies it must be that they do not share at least one element y_j from one of the elements $y_0, y_1, \ldots, y_{K-1}$ (otherwise they are the same vector). But then there exists a vector $u \in \mathcal{U}$ such that u is the same as vector $u_{\ell+1}$ except with the last element of $u_{\ell+1}$ replaced with y_{2K}. Then $u \notin A'$ and $G(u) \subset G(A')$, so that A' is redundant.

2. $u_{\ell+1}$ and u_1 share at least one of the elements $y_0, y_1, \ldots, y_{K-1}$: Note that if these elements share y_{1K} then we are in the previous case, since this implies that they do not share at least one element in $y_0, y_1, \ldots, y_{K-1}$. Thus, suppose they do not share y_{1K}. If they share all other elements, then $u_{\ell+1}$ shares exactly $K-1$ elements with all vectors in A, which is a contradiction. Thus, there must exist at least one element in $y_0, y_1, \ldots, y_{K-1}$ that they do not share. But note there exists a $u \in \mathcal{U}$ that is the same as u_1 except that its last element is replaced with the last element of $u_{\ell+1}$. But then $u \notin A'$ and $G(u) \subset G(A')$, so that A' is redundant.

We conclude that $u_{\ell+1}$ must have the same elements in common with $u_1, u_2, \ldots, u_\ell$, which shows the inductive step and concludes the proof. ■

Proof of Lemma 2. For notational simplicity, let $M \equiv |\mathcal{Y}|$ and $K \equiv |\mathcal{D}|$.

(a) First note that for any $(y, d), (y', d) \in \mathcal{W}$ we have $G^{-1}(y, d) \cap G^{-1}(y', d) = \emptyset$. Thus we can divide the graph $\mathcal{G}$ into K disjoint subgraphs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K$ where $\mathcal{G}_k = (\mathcal{W}_k, \mathcal{U}, G)$ and where

$$\mathcal{W}_k = \{(y, d) : d = k\}$$

By construction we have $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ for all $i \neq j$, and $G^{-1}(y,d) \cap G^{-1}(y',d) = \emptyset$ for any $y \neq y'$. Also note that the vectors of the form (y,d) map to vectors of the form $(\cdot,\cdot,\ldots,\cdot,y,\cdot,\ldots,\cdot)$, with y in the d^{th} position. Thus, collecting such vectors for all values of y we obtain the collection $\mathcal{U}$, so that we can conclude $G^{-1}(\mathcal{W}_k) = \mathcal{U}$. Finally consider the pair (v,v') with $v \in \mathcal{W}_i$, $v' \in \mathcal{W}_j$, $i \neq j$. v and v' can be written as $v = (y,i)$ and $v' = (y',j)$. But since v is mapped to the set of vectors of the form $(\cdot,\cdot,\ldots,\cdot,y,\cdot,\ldots,\cdot)$, with y in the i^{th} position, and since v' is mapped to the set of vectors of the form $(\cdot,\cdot,\ldots,\cdot,y',\cdot,\ldots,\cdot)$, with y' in the j^{th} position, it is clear that $G^{-1}(v) \cap G^{-1}(v') \neq \emptyset$ when $i \neq j$.

(b) For the forward direction note that by property (iii) of collections $\mathcal{W}_k$ proved in part (a), (i) is implied if B is self-connected. In addition, note that $G^{-1}(\mathcal{W}_k) = \mathcal{U}$ for every k, so that if (ii) did not hold for $B \in \mathcal{S}_w$ we would have $G^{-1}(B) = \mathcal{U}$. But then if $B \neq \mathcal{W}$ we can always find a $v \notin B$ such that $G^{-1}(v) \subset G^{-1}(B)$, contradicting the fact that $B \in \mathcal{S}_w$.

For the reverse, note first that since $G^{-1}(y,d) \cap G^{-1}(y',d) = \emptyset$ for any $y \neq y'$, and $G^{-1}(y,d) \cap G^{-1}(y',d') \neq \emptyset$ for any $d \neq d'$, condition (i) is sufficient to ensure B is self-connected. Next, suppose by way of contradiction that there exists a collection of singletons $B = \{y_1,\ldots,y_r\} \subset \mathcal{W}$ satisfying conditions (i) and (ii), but that there also exists a $v \in \mathcal{W}$ such that $v \notin B$ and $G^{-1}(v) \subset G^{-1}(B)$. Note that v can be written as $v = (y,d)$, and maps to the set of vectors of the form $(\cdot,\cdot,\ldots,\cdot,y,\cdot,\ldots,\cdot)$, with y in the d^{th} position. Thus $G^{-1}(B)$ must contain all the vectors of this form if $G^{-1}(v) \subset G^{-1}(B)$. But since B does not contain v, this is only possible if $\mathcal{W}_k \subseteq B$ for some k, contradicting the fact that condition (ii) is satisfied.

■

Proof of Lemma 3. For notational simplicity, let $M \equiv |\mathcal{Y}|$ and $K \equiv |\mathcal{D}|$. Consider any $A \in \mathcal{S}_u$ with $|A| = r$. We want to show there exists a $B \in \mathcal{S}_w$ such that $A = G^{-1}(B)^c$, or equivalently, $A^c = G^{-1}(B)$. Since $A \in \mathcal{S}_u$, by Lemma 1 the singletons that comprise A have exactly $K-1$ elements in common. Suppose without loss of generality that the uncommon element is the first element, and suppose the $K-1$ common elements are $y_1, y_1, \ldots, y_1$. Then every $u_i \in A$ can be written

$$u_i = (v_i, y_1, y_1, \ldots, y_1)$$

for some $v_i \in \{y_1, y_2, \ldots, y_M\}$, and where $v_i \neq v_j$ for $i \neq j$. Given our $A \in \mathcal{S}_u$ described above, A^c can be represented by

$$A^c = \{\{u_i\}_{i=1}^r : u_i = (v_i, y_1, y_1, \ldots, y_1),\ v_i \in \{y_1, y_2, \ldots, y_M\},\ i = 1, \ldots, r\}^c$$

$$= \left(\bigcup_{i_1=r+1}^{M} \bigcup_{i_2=1}^{M} \bigcup_{i_3=1}^{M} \cdots \bigcup_{i_K=1}^{M} (v_{i_1}, y_{i_2}, y_{i_3}, \ldots, y_{i_K}) \right)$$
$$\cup \left(\bigcup_{i_1=1}^{M} \bigcup_{i_2=2}^{M} \bigcup_{i_3=1}^{M} \cdots \bigcup_{i_K=1}^{M} (v_{i_1}, y_{i_2}, y_{i_3}, \ldots, y_{i_K}) \right) \cup \ldots$$
$$\ldots \cup \left(\bigcup_{i_1=1}^{M} \bigcup_{i_2=1}^{M} \bigcup_{i_3=1}^{M} \cdots \bigcup_{i_K=2}^{M} (v_{i_1}, y_{i_2}, y_{i_3}, \ldots, y_{i_K}) \right)$$

$$= \left(\bigcup_{i_1=r+1}^{M} G^{-1}(v_{i_1}, 1) \right) \cup \left(\bigcup_{j=2}^{M} \bigcup_{k=2}^{K} G^{-1}(y_j, k) \right)$$

$$= G^{-1} \left(\bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (v_{i_1}, 1) \cup (y_j, k) \right)$$

Now set

$$B = \bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (v_{i_1}, 1) \cup (y_j, k)$$

and consider the follow cases:

- $M > K, K = 2$: We claim $B \in \mathcal{S}_w$ only if $1 \leq r \leq M-1$. Indeed, if $r \geq |\mathcal{Y}|$ then

$$B = \bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (v_{i_1}, 1) \cup (y_j, k) = \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (y_j, k) = \bigcup_{j=2}^{M} (y_j, 2)$$

  so that clearly $B \subseteq \mathcal{W}_2$ and so $B \notin \mathcal{S}_w$. However, if $1 \leq r \leq M-1$ then

$$B = \bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (v_{i_1}, 1) \cup (y_j, k) = \bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} (v_{i_1}, 1) \cup (y_j, 2)$$

  so $B \not\subseteq \mathcal{W}_k$ for any k and $\mathcal{W}_k \not\subseteq B$ for any k, which proves $B \in \mathcal{S}_w$ by Lemma 2.

- $K \geq 3$: We claim that $B \in \mathcal{S}_w$ with no additional conditions. This follows from the fact that the union:

$$\bigcup_{i_1=r+1}^{M} \bigcup_{j=2}^{M} \bigcup_{k=2}^{K} (v_{i_1}, 1) \cup (y_j, k)$$

  contains elements from $\mathcal{W}_2, \ldots, \mathcal{W}_k$ regardless of the magnitude of r, and $\mathcal{W}_k \not\subseteq B$ for any k. Thus by Lemma 2 we have that $B \in \mathcal{S}_w$.

Thus we conclude that if $K = 2$ and $K < M$, then for any $A \in \mathcal{S}_w$ with $|A| \leq M-1$, $\exists B \in \mathcal{S}_w$ such that $A^c = G^{-1}(B)$, so that $A \in \mathcal{S}_w^{-1}$. Otherwise, if $K > 2$, then for any $A \in \mathcal{S}_w$, $\exists B \in \mathcal{S}_w$ such that $A^c = G^{-1}(B)$, so that $A \in \mathcal{S}_w^{-1}$. This completes the proof. ■

Proof of Theorem 1. For notational simplicity, let $M \equiv |\mathcal{Y}|$ and $K \equiv |\mathcal{D}|$.

1. Note that every singleton trivially satisfies the conditions in Definition 5, so that the result holds for $r = 1$. Now consider any $A \in \mathcal{S}_u$ with $|A| = r \geq 2$. We know from Lemma 1 that every $u \in A$ must share the same $K-1$ elements. There are M^{K-1} ways

to select the first $K-1$ elements, and $\binom{M}{r}$ ways of choosing the uncommon element. Finally, the uncommon element can be in any one of K positions. We conclude that there are exactly

$$M^{K-1}K \cdot \binom{M}{r}$$

sets $A \in \mathcal{S}_u$ with $|A| = r \geq 2$.

2. By the results of Lemma 2, to construct a set $B \in \mathcal{S}_w$ of size r from the singletons we can choose r elements from any combination of the K subsets $\mathcal{W}_k$, but we must choose elements from at least two subsets, and we must choose less than M elements from each collection. Now note that there are $\binom{K}{\ell}$ ways to choose from any $2 \leq \ell \leq K$ collections, and $\binom{M}{v_k}$ ways to choose $1 \leq v_k \leq M-1$ elements from each collection. Finally, we must ensure that if we are constructing an r-element set B that we have

$$\sum_k v_k = r$$

Combining everything, there are

$$\sum_{\ell=2}^{K} \binom{K}{\ell} \left(\sum_{v \in A(r,M,\ell)} \prod_{i=1}^{\ell} \binom{M}{v_i} \right)$$

r-element sets in the collection $\mathcal{S}_w$, where

$$A(r, M, \ell) = \left\{ (v_1, v_2, \ldots, v_\ell) \in \mathbb{N}^\ell : \sum_i v_i = r, \quad 1 \leq v_i \leq M-1 \; \forall i \right\}$$

as claimed.

3. This follows from part 1 of this Theorem when combined with Lemma 3.

■

Proof of Theorem 2. Notation for the proof is given in Appendix 4.

By Theorem 1 the identified set Θ^f is an interval. Thus, to show consistency with respect to the Hausdorff metric, it suffices to show that $\hat{f}_n^\ell \xrightarrow{p} f^\ell$ and $\hat{f}_n^u \xrightarrow{p} f^u$. We can focus on the

upper bound problem, since the lower bound problem is symmetric. The upper bounding problem is:

$$f^u(\mathbb{P}_n) = \sup_{Q \in \mathcal{Q}(\mathbb{P}_n)} f(Q) \tag{7}$$

To prove consistency we want to show that for every $\varepsilon > 0$:

$$\limsup_{n \to \infty} P\left(|f^u(\mathbb{P}_n) - f^u(P)| > \varepsilon\right) = 0 \tag{8}$$

Now note:

$$|f^u(\mathbb{P}_n) - f^u(P)| = \left| \sup_{Q \in \mathcal{Q}(\mathbb{P}_n)} f(Q) - \sup_{Q \in \mathcal{Q}(P)} f(Q) \right| \leq \sup_{||Q - Q'|| \leq d_H(\mathcal{Q}(\mathbb{P}_n), \mathcal{Q}(P))} |f(Q) - f(Q')|$$

Let Δ_Q denote the $(|\mathcal{U}| - 1)$-simplex. Since $\Delta_Q \subset \mathbb{R}^{d_\mathcal{U}}$ is compact, continuity of f implies uniformly continuity over Δ_Q. Thus, we know that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $||Q - Q'|| < \delta$ implies $|f(Q) - f(Q')| < \varepsilon$. Thus, to show (8) it suffices to show that for every $\delta > 0$:

$$\limsup_{n \to \infty} P\left(d_H(\mathcal{Q}(\mathbb{P}_n), \mathcal{Q}(P)) > \delta\right) = 0 \tag{9}$$

Note that by assumption (a) (and the fact Artstein's Theorem implies only linear inequality constraints) $\mathcal{Q}(\cdot)$ is defined completely by linear equality and inequality constraints. Now convert all inequality constraints to equality constraints by introducing a slackness parameter $\lambda_k \geq 0$ for each constraint. Let λ denote the vector of slackness parameters, and let $\theta = (Q', \lambda')'$ be the vector of dimension $d_\theta \times 1$. In addition, let $g(\theta, P)$ be the $d_e \times 1$ vector of moment equalities. Rather than include the constraint $\sum_{u \in \mathcal{U}} Q(U = u) = 1$ as an equality constraint, note that, as per the remark 1 in Shi and Shum (2015), we can instead drop one equality constraint $g_k(\theta, P)$ (and thus also the associated slackness parameter λ_k), and solve for λ_k using the constraint:

$$\sum_{j \in I(\mathcal{U})} Q(U = u_j) + \sum_{j \in I(\mathcal{U})} \lambda_j = 1$$

$$\implies \lambda_k = 1 - \sum_{j \in I(\mathcal{U})} Q(U = u_j) - \sum_{j \in I(\mathcal{U}), j \neq k} \lambda_j \tag{10}$$

and then add the non-negativity constraint on (10) (where $I(\mathcal{U})$ is an index set for elements in $\mathcal{U}$). Thus, there will be $(d_e - 1)$ equality constraints in the vector $g(\theta, P)$, and d_θ inequality constraints given by the vector

$$h(\theta) \equiv \begin{pmatrix} Q \\ \lambda_{-k} \\ 1 - \sum_{j \in I(\mathcal{U})} Q(U = u_j) - \sum_{j \in I(\mathcal{U}), j \neq k} \lambda_j \end{pmatrix} \succcurlyeq \mathbf{0}$$

Importantly, note that the inequality constraints do not depend on the first-stage parameter P. Now define $\Theta(P) = \{\theta \in \Theta : g(\theta, P) = 0, h(\theta) \geq 0\}$. Consider the criterion function:

$$T(\theta, P) = g(\theta, P)' g(\theta, P)$$

Then under assumption (d) we have:

$$\Theta(P) = \arg\min_{\theta \in \Theta} T(\theta, P) \quad s.t. \quad h(\theta) \succcurlyeq \mathbf{0}$$

The sample analog of the above is:

$$\Theta(\mathbb{P}_n) = \arg\min_{\theta \in \Theta} T(\theta, \mathbb{P}_n) \quad s.t. \quad h(\theta) \succcurlyeq \mathbf{0}$$

Under assumption (d), $d_H(\mathcal{Q}(\mathbb{P}_n), \mathcal{Q}(P)) \xrightarrow{p} 0$ if and only if $d_H(\Theta(\mathbb{P}_n), \Theta(P)) \xrightarrow{p} 0$. Thus it suffices to show the latter. To do this, we will verify the conditions of Theorem 2.1 in Shi and Shum (2015):

1. Since $2^{\mathcal{W}}$ contains at most a finite number of sets, by assumption (c) and the Glivenko-Cantelli Theorem we know that $\sup_{A \in 2^{\mathcal{W}}} |\mathbb{P}_n(A) - P(A)| = o_P(1)$; thus, $\mathbb{P}_n$ converges uniformly to P in probability.

2. The $(|\mathcal{W}| - 1)$-simplex $\Delta_P \subset \mathbb{R}^{d_{\mathcal{W}}}$ is compact. Θ is also compact (since it is without

loss of generality that we restrict $\lambda \in [0, 1]$).

3. $g(\cdot, P)$ is trivially continuously differentiable on Θ for all P, and $h(\cdot)$ is trivially continuous on Θ; this follows since both $g(\cdot, P)$ and $h(\cdot)$ are linear functions of θ.

4. Note by assumption (a) that $\Theta(P)$ is defined completely by linear equality and inequality constraints and is closed and convex, so that together with assumption (d) we have $cl(int(\Theta(P))) = \Theta(P)$ (see Remark (i) after Theorem 2.1 in Shi and Shum (2015)). In addition, by assumption (b) the Jacobian $\partial g(\theta, P)/\partial \theta'$ must have full row rank. To see this, first note by linearity of all constraints the Jacobian is a matrix of constants. Next note all equality constraints can be classified as (i) equality constraints defining $\mathcal{Q}^{\dagger}$, and (ii) equality constraints that were converted from inequality constraints by adding a slackness parameter. By assumption (b), the Jacobian of the set of linear equality constraints of type (i) have full row rank. For equality constraints of type (ii) the rows will also have full rank, since by construction any equality constraint j that was constructed from an inequality constraint will contain its own slackness parameter λ_j (and thus row j contains a 1 in the Jacobian for λ_j, and row $j' \neq j$ contains a 0 for λ_j). Finally, note that equality constraints of type (ii) can be combined with equality constraints of type (i) while still yielding a full rank Jacobian. This last step again follows since type (ii) equality constraints will contain additional non-zero entries in the rows of the Jacobian for the slackness parameters, so that the gradients of these constraints will not be linearly dependent with the gradients of the constraints of type (i), which do not contain such non-zero entries.

Consistency of $\Theta(\mathbb{P}_n)$ for $\Theta(P)$ in the Hausdorff metric then follows from Theorem 2.1 in Shi and Shum (2015).

■

References

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

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

Demuynck, T. (2015). Bounding average treatment effects: A linear programming approach. *Economics Letters*, 137:75–77.

Lafférs, L. (2013). *Essays in partial identification*. PhD thesis, Department of Economics, NHH-Norwegian School of Economics.

Lafférs, L. (2015). Bounding average treatment effects using linear programming. Technical report, CEMMAP working paper, Centre for Microdata Methods and Practice.

Luo, Y. and Wang, H. (2016). Core determining class: Construction approximation and inference. Working paper.

Luo, Y. and Wang, H. (2017). Core determining class and inequality selection. *American Economic Review Papers and Proceedings*, 107(5):274–277.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68(4):997–1010.

Menzel, K. (2014). Consistent estimation with many moment inequalities. *Journal of Econometrics*, 182(2):329–350.

Molchanov, I. (2005). *Theory of random sets*. Springer Science & Business Media.

Shi, X. and Shum, M. (2015). Simple two-stage inference for a class of partially identified models. *Econometric Theory*, 31(3):493–520.

Torgovitsky, A. (2016). Nonparametric inference on state dependence with applications to employment dynamics. Working paper.

Yildiz, N. (2012). Consistency of plug-in estimators of upper contour and level sets. *Econometric Theory*, 28(2):309–327.