Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects

Thomas M. Russell*

University of Toronto

December 15, 2017

Abstract

This paper investigates identification and estimation of bounds on continuous functionals of the joint distribution of potential outcomes in the program evaluation literature. As discussed in Heckman, Smith and Clements (1997), there are many interesting parameters in the program evaluation literature that require knowledge of the dependence between potential outcomes across treated and untreated states. However, the joint distribution of potential outcomes is a complicated object, and it can be difficult to construct bounds on lower-dimensional functionals of the joint distribution, as well as to verify that the proposed bounds are sharp. This paper proposes an identification and estimation method that allows researchers to easily bound continuous functionals of the joint distribution in a completely nonparametric setting without the need to verify sharpness on a case-by-case basis. The focus is on a model where the selection mechanism is left completely unspecified. The method can sharply bound interesting parameters with analytic bounds that may be difficult to derive, can be used in settings in which instruments are available, and can easily accommodate additional model constraints. However, computational considerations for the method are found to be important, and are discussed in detail. Supplementary materials for this article are available online.

JEL classification: C01, C13, C14, C31.

Keywords: Partial Identification, Linear Programming, Convex Optimization, Random Set Theory, Optimal Transport

1 Introduction

This paper investigates identification and estimation of bounds on continuous functionals of the joint distribution of potential outcomes in the program evaluation literature. The focus throughout is on a

I am indebted to Ismael Mourifié for countless hours of discussion that have helped to improve this paper. I am also especially grateful to Victor Aguirregabiria, JoonHwan Cho, Christian Gourieroux, Jiaying Gu, and Yuanyuan Wan who have all provided excellent feedback, and to Marc Henry for hosting me at the Penn State Department of Economics. The paper benefited from conversations with Stephane Bonhomme, Sung Jae Jun, Desiré Kedagni, Toru Kitagawa, Lixiong Li, Joris Pinske, and Adam Rosen. All errors are my own.

*Thomas Russell, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, M5S3G7, Canada. Email: thomas.russell@mail.utoronto.ca.

general version of the potential outcome model commonly used in the analysis of treatment effects. Here we consider the case where the mechanism governing the treatment decision is left completely unrestricted. In cases where the selection mechanism is left unspecified it is still largely unknown how to obtain a tractable characterization of the identified set for a general class of parameters, for arbitrary discrete-valued outcomes and treatments, in a completely nonparametric setting. Indeed, there are still many important parameters in the program evaluation literature, some very simple, for which no closed-form bounds exist. The fact that no closed-form bounds exist is often because bounding such parameters requires knowledge of the dependence of potential outcomes across treated and untreated states; i.e. knowledge of the joint distribution of potential outcomes. This point has been appreciated by Heckman et al. (1997), who argue that there are many parameters useful to policy makers that require knowledge of the joint distribution. The procedure proposed in this paper allows researchers to bound most of the parameters proposed by Heckman et al. (1997), and more, in a completely nonparametric setting. Some examples of parameters that can be written as continuous functionals of the joint distribution include the average treatment effect, the correlation between potential outcomes, conditional probabilities, and the variance of treatment effects. Examples of situations where these parameters are interesting will be discussed. The method is also amenable to the sequential introduction of additional constraints on the identified set and the inclusion of instruments, which should be appealing to applied researchers. In addition, although we consider the case where the selection mechanism is left unspecified, the framework can also be used in models that impose more structure. As such, many of the bounds previously proposed in the literature can be obtained as a special case. Finally, unlike in the recent paper of Fan et al. (2017), the method does not require that the marginal distributions of potential outcomes be identified.

The method will accomplish these goals by abandoning the analytic approach to characterizing the identified set. As seen in Mourifie et al. (2015), closed-form expressions for bounds on functions of the joint distribution can be difficult to derive. In addition, even when meaningful bounds can be proposed, proving the bounds are sharp can be a delicate operation. Instead, we do not derive closed-form analytic expressions for the bounds on any parameter of interest, but rather show that the bounding problem can solved as a minimization (for the lower bound) and maximization (for the upper bound) problem subject to a carefully selected set of constraints. To bound a continuous functional of the joint distribution in this way requires that the constraints in the optimization problem reflect all of the restrictions imposed on the joint distribution by the observed distribution. We compare two characterizations of the complete set of restrictions imposed on the joint distribution of potential outcomes in terms of their computational tractability. The first characterization is based on *Artstein's Theorem* (Artstein (1983)) from random set theory. This characterization has been explored previously by Galichon and Henry (2011), Beresteanu et al. (2012) and Chesher and Rosen (2017), among others, and is also used in the main identification results for this paper. The second characterization re-frames the bounding problem as an optimal transport problem, and has been considered in Galichon and Henry (2011) and Laffërs (2013b, 2015). The two characterizations are compared based on their computational tractability, and the conditions under which one approach dominates the other are discussed.

Finally, we apply the theoretical results to data from the Tennessee STAR experiment considered in Krueger (1999) and Krueger and Whitmore (2001). We find that bounds on the average treatment are informative and consistent with the results of Krueger (1999). However, we also find informative bounds on parameters such as the correlation between potential outcomes —measuring dependence across counterfactual states— and the standard deviation of treatment effects —measuring the heterogeneity of treatment effects. The application shows how bounds on a battery of parameters can be useful in constructing a complete picture of the effects of the program, and also demonstrates the sensitivity of identification to modeling assumptions.

This paper is an extension of work by Galichon and Henry (2006, 2009, 2011), Beresteanu and Molinari (2008), Beresteanu et al. (2011, 2012), and more recently Chesher and Rosen (2017). Similar to Mourifie et al. (2015), we construct bounds without imposing structure on the selection mechanism. This approach follows the philosophy of Manski (2003, 2009) who suggests that researchers first ask what can be learned from the data alone before imposing additional assumptions. When credible assumptions are available, the procedure in this paper also serves as a framework to facilitate the introduction of additional model assumptions and structure.[1]

This study is also similar in spirit to growing work in computational approaches to partial identification. Early work in this literature was done by Balke and Pearl (1994) for bounds on counterfactual probabilities. Bounds on the average treatment effect under a variety of assumptions using linear programming are presented in Chiburis (2010), Laffers (2015), Demuynck (2015), and Torgovitsky (2016). Outside of treatment effects, other interesting uses of linear programming in partial identification can be found in Honoré and Lleras-Muney (2006), Honoré and Tamer (2006), Manski (2007) and Molinari (2008).

This remainder of the paper is organized as follows. Section 2 describes the general bounding problem and an identification result for bounding functionals of the joint distribution. Section 3 presents two practical approaches to estimation based on the identification result. Section 4 presents the application to data from the Tennessee STAR experiment, and section 5 concludes. A variety of results are available in the online supplementary materials, including proofs of the main results.

2 Identification

2.1 Preliminaries

Recall that in typical treatment effect models we observe realizations of the random variables $(Y, D) \in \mathcal{Y} \times \mathcal{D}$, where Y represents the outcome variable, and D represents the finite-valued treatment variable. Without

[1] The approach is also in the spirit of Ginther (2000), who shows estimates of returns to school to the selection mechanism specified by the researcher. If results are sensitive to the imposed selection mechanism, then remaining agnostic on the nature of selection may be the only credible approach.

loss of generality we take $\mathcal{D} = \{0, 1, \ldots, K-1\}$. When possible, we will use the notation $W \equiv (Y, D)$ and $\mathcal{W} \equiv \mathcal{Y} \times \mathcal{D}$. With this defined, we assume throughout that $\mathcal{W}$ is a finite subset of euclidean space, and denote the σ-algebra on $\mathcal{W}$ as $2^{\mathcal{W}}$. In treatment effect models there is also an unobserved random vector $U \equiv (Y_0, Y_1, \ldots, Y_{K-1}) \in \mathcal{U}$, where we assume that $\mathcal{U} = \mathcal{Y}^K$; i.e. the support of each variable Y_d for $d = 0, \ldots, K-1$ is common and equal to the finite support $\mathcal{Y}$ of Y.[2] Finally, we denote the $\sigma-$algebra on U as $2^{\mathcal{U}}$ and we will refer to the vector of random variables U as *potential outcomes.*

All random variables in this paper are assumed to be defined on the same underlying probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Let P denote the distribution induced on $\mathcal{W}$ by (Y, D), and let Q denote the distribution induced by U on $\mathcal{U}$. In particular:

$$\begin{aligned} P(A) &= \mathbb{P}(W \in A) \qquad \forall A \in 2^{\mathcal{W}}, \\ Q(B) &= \mathbb{P}(U \in B) \qquad \forall B \in 2^{\mathcal{U}}. \end{aligned}$$

Combining everything leads to the following familiar definition of the *potential outcome model*:

Definition 1 (Potential Outcome Model)**.** *A Potential Outcome Model (POM) is one in which Y is determined by:*

$$Y = \sum_{d=0}^{K-1} Y_d \mathbb{1}\{D = d\},$$

where $|\mathcal{Y}| \geq 2$ and $K \geq 2$.

The objective is to recover the distribution Q, and functionals thereof, using only knowledge of the distribution of the observed random variables (Y, D). For now we assume that the researcher has access to an infinite independent and identically distributed sample $\{Y_i, D_i\}_{i=1}^{\infty}$ drawn from the distribution P. Issues of sampling uncertainty are addressed in Appendix D in the online supplementary material.

The fundamental problem of causal inference is that we do not observe a full realization of the vector $(Y_0, Y_1, \ldots, Y_{K-1})$ for any individual. In addition, in the absence of randomly-assigned treatment, there may be dependence between the random variables D and U. Because of these issues, even simple parameters such as the *average treatment effect* are impossible to point-identify without additional assumptions.

Because of the possible dependence between potential outcomes U and the treatment status D, researchers typically introduce an *instrumental variable* $Z : \Omega \to \mathcal{Z}$, where we assume that $\mathcal{Z}$ is a finite set. The instrument Z is a random variable that is assumed to affect the treatment choice D but to be independent of potential outcomes U (denoted by $Z \perp\!\!\!\perp U$). Since the use of an instrument is common in the program evaluation literature, we extend all results for the POM to the case where an instrument is available. In such a case, we will let P_z denote the conditional distribution induced on $\mathcal{W}$ by (Y, D) given $Z = z$. In particular:

$$P_z(A) = \mathbb{P}(W \in A | Z = z) \qquad \forall A \in 2^{\mathcal{W}}.$$

[2] This assumption means that the support of the random variable U is informed by the support of the observed outcomes; although natural, researchers may find this restrictive in some circumstances.

In this paper we leave the dependence between Z and D completely unrestricted. Indeed, these variables may be highly dependent, or completely independent. In addition, since this paper deals with partial identification in a nonparametric setting, the instrument Z will not be needed to solve the endogeneity problem that typically arises in parametric models of the dependence between D and U. Instead, as we will see, the instrument Z will only be used to produce more informative bounds on the parameter of interest.

With the preliminaries in place, we provide some discussion of functionals f of the joint distribution Q. Heckman et al. (1997) argue that there are many interesting parameters in the program evaluation literature that require knowledge of the joint distribution Q. One trivial example is to take f to represent the average treatment effect:

$$f(Q) = \int_{\mathcal{U}} (Y_1 - Y_0)\, dQ.$$

However, we might also consider other less typical parameters (in no particular order):

(i) *The correlation between potential outcomes:*

$$f(Q) = \frac{\int_{\mathcal{U}} (Y_0 - \mathbb{E}(Y_0))\,(Y_1 - \mathbb{E}(Y_1))\, dQ}{\left(\int_{\mathcal{U}} (Y_0 - \mathbb{E}(Y_0))^2\, dQ\right)^{1/2} \left(\int_{\mathcal{U}} (Y_1 - \mathbb{E}(Y_1))^2\, dQ\right)^{1/2}}.$$

This parameter provides a simple measure of the dependence of outcomes across the treated and untreated states. The importance of capturing the dependence across counterfactual states is well-illustrated in Honoré and Lleras-Muney (2006) in a competing risk model of cancer and cardiovascular disease. It is also well-illustrated in the application of Mourifie et al. (2015) to the case of the STEM versus non-STEM field choice, where the level of dependence across counterfactual states may determine policy recommendations.[3] Note that to bound the correlation coefficient, one must jointly bound the mean and variance of potential outcomes: in general one cannot recover sharp bounds on the correlation coefficient by first bounding the mean and variance Y_0 and Y_1, and then computing bounds for the correlation coefficient via a 'plug-in' estimator. Given this difficulty, it is not clear how one might bound this parameter analytically.

(ii) *Voting Criterion:*

$$f(Q) = \int_{\mathcal{U}} \mathbb{1}\{Y_1 > Y_0\} dQ.$$

This parameter provides a measure of the proportion of individuals who benefit from treatment. This parameter is discussed in Heckman and Vytlacil (2007) as an important parameter that requires knowledge of the joint distribution. Closed-form bounds for this parameter in the binary outcome case are provided by Mourifie et al. (2015).

[3] STEM stands for Science, Technology, Engineering and Mathematics.

(iii) *Distributional Mobility:*

$$f(Q) = Q(Y_1 \in A | Y_0 \in B) = \frac{\int \mathbb{1}\{Y_1 \in A, Y_0 \in B\} dQ}{\int \mathbb{1}\{Y_0 \in B\} dQ}.$$

This parameter measures the probability that treatment helps an individual obtain an outcome $Y_1 \in A$, given his/her untreated outcome is fixed at $Y_0 \in B$. Mourifie et al. (2015) provide bounds for this parameter, but do not claim sharpness in the presence of an instrument. Indeed, in the presence of an instrument, no known closed-form sharp bounds exist for this parameter.

(iv) *Variance of Treatment Effects:*

$$f(Q) = \int_{\mathcal{U}} \left((Y_1 - Y_0) - \mathbb{E}(Y_1 - Y_0) \right)^2 dQ.$$

The variance of treatment effects can provide a measure of the heterogeneity of treatment effects. If the potential outcomes Y_1 and Y_0 are dependent, then this parameter will require knowledge of the joint distribution Q.

Note that not every parameter is a continuous functional of the joint distribution. For example, the *interquartile range*, for which sharp bounds are provided by Mourifie et al. (2015), in general cannot be expressed as a continuous functional of the joint distribution.

The remainder of this section describes a general framework that can be used to derive sharp bounds on any parameter that can be written as a continuous functional of the joint distribution of potential outcomes. The method is based on the following intuition. First, we characterize the set of all distributions Q that are consistent with the observed distribution P and the researcher's assumptions. Denote this set as $\mathcal{Q}$. Under the assumption that $\mathcal{U}$ is finite, this set will be convex and compact with respect to the usual topology. Next, we will bound any continuous function $f : \mathcal{Q} \to \mathbb{R}$ by noting that the image of a continuous function over a compact set is an interval $[f^\ell, f^u]$ where:

$$f^u = \sup_{Q \in \mathcal{Q}} f(Q) \qquad\qquad f^\ell = \inf_{Q \in \mathcal{Q}} f(Q). \tag{1}$$

Obtaining sharp bounds on the function f then reduces to solving these two optimization problems.

2.2 Identification without an Instrument

The formal identification argument will use results from random set theory.[4] Before beginning, we define the model as $(G, \mathcal{Q}^\dagger)$, where $\mathcal{Q}^\dagger$ is the collection of all admissible distributions and G is the model correspondence $G : \mathcal{U} \to \mathcal{Y}$. The set of admissible distributions $\mathcal{Q}^\dagger$ represents the set of all distributions that satisfy our *a*

[4] Random set theory is a convenient tool in partial identification, and has been used previously by Galichon and Henry (2006), Galichon and Henry (2009), Galichon and Henry (2011), Beresteanu and Molinari (2008), Beresteanu et al. (2011, 2012), and Chesher and Rosen (2017), among others.

priori restrictions on the distribution of U.[5] In the absence of restrictions, we can simply take $\mathcal{Q}^\dagger$ to be the K-dimensional probability simplex. The model correspondence for the POM is given by:

$$G(y_0, y_1, \ldots, y_{K-1}) = \left\{ (y, d) \in \mathcal{W} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d = k\} \right\}, \tag{2}$$

for $(y_0, y_1, \ldots, y_{K-1}, z) \in \mathcal{U}$. This model is *incomplete* in the sense that the process generating the (possibly endogenous) variable D has not been specified. In incomplete models the outcome variables cannot be uniquely determined by the values of the latent variables, and the mapping $G : \mathcal{U} \to \mathcal{W}$ is given by a correspondence rather than a function.[6] There are two immediate advantages of modeling in this way. First, any unobserved heterogeneity affecting choices and/or outcomes is left completely unrestricted.[7] Second, although there may be weak assumptions on the selection mechanism that "complete" the incomplete POM described here, complete models are a special case of incomplete models, but not the reverse. Thus, while we focus (for now) on what can be identified from the data alone, "completing" the model with additional assumptions can be accommodated by the framework discussed in this paper under minor modifications.[8]

Taking the incomplete nature of this model as given, it is equivalent to focus on learning the distribution Q through the reverse correspondence:

$$G^{-1}(y, d) = \left\{ (y_0, y_1, \ldots, y_{K-1}) \in \mathcal{U} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d = k\} \right\}. \tag{3}$$

Replacing the arguments (y, d) of the correspondence G^{-1} with the random variables (Y, D), we have that $G^{-1}(Y, D) : \Omega \to 2^{\mathcal{U}}$ is a *random closed set* (see Appendix A in the online supplementary material for details). There are possibly many random variables $\widetilde{U}$ that can map within this random set. We say that a given random variable $\widetilde{U}$ can rationalize the distribution of (Y, D) if $\widetilde{U} \in G^{-1}(Y, D)$ a.s. The incomplete nature of the POM implies that there are many random variables $\widetilde{U}$ —and thus many distributions induced by the latent variables $\widetilde{U}$— that could rationalize the observed distribution $P(Y, D)$ given the correspondence in (2). In such models, the observed distribution characterizes the random set $G^{-1}(Y, D)$ through the *generalized likelihood*:

$$T(A) \equiv P((Y, D) : G^{-1}(Y, D) \cap A \neq \emptyset), \tag{4}$$

defined for every $A \in 2^{\mathcal{U}}$; see the discussion in Galichon and Henry (2011). The functional T is sometimes

[5] For example, we may take $\mathcal{Q}^\dagger$ to be the set of distributions that satisfy the *monotone treatment response* or *monotone treatment selection* conditions discussed in Manski and Pepper (2000). These assumptions are discussed further in Appendix C of the online supplementary material. Alternatively, in the presence of an instrument, we might consider the independence, mean independence, and quantile independence conditions discussed in Chesher and Rosen (2017).

[6] Defined analogously, a complete model is one in which the outcome variables are uniquely determined by the latent variables. This definition is consistent with that in Jovanovic (1989). For an especially clear discussion of the distinction with complete models, see Chesher and Rosen (2012).

[7] For an example of a case where this matters, consider the results of Ginther (2000) who shows the sensitivity of estimates of the returns to schooling to assumptions on the selection mechanism.

[8] In this sense we follow the philosophy of Manski (2003, 2009) by first providing researchers a means of computing bounds under minimal assumptions. After computing these bounds as a first-pass, researchers may then impose credible assumptions to increase the informativeness of the analysis.

called the *capacity functional* of the *random set* $G^{-1}(Y, D)$ (see Appendix A in the online supplementary material for a formal definition). Note that, given $G : \mathcal{U} \to 2^{\mathcal{W}}$ is also a random set, we could have also defined a capacity functional for the random set $G(U)$ as:

$$T_{\mathcal{W}}(A) \equiv P(U : G(U) \cap A \neq \emptyset), \tag{5}$$

for every compact $A \in 2^{\mathcal{W}}$. We say that the capacity functional given in (4) has been defined on the *observables*, whereas the capacity functional given in (5) has been defined on the *unobservables*. Although we focus on the case where the capacity functional is defined on the observables, from an identification perspective, these characterizations are equivalent; see Chesher and Rosen (2017) for a discussion.

Example 1. Consider the POM and suppose we are in a binary outcome, binary treatment setting, where $Y \in \{0, 1\}$ and $D \in \{0, 1\}$. Then we will have:

$$\mathcal{Y} = \{(Y, D) : Y \in \{0, 1\}, D \in \{0, 1\}\} = \{(0, 0), (1, 0), (0, 1), (1, 1)\},$$

$$\mathcal{U} = \{(Y_0, Y_1) : Y_0 \in \{0, 1\}, Y_1 \in \{0, 1\}\} = \{(0, 0), (1, 0), (0, 1), (1, 1)\}.$$

In addition, we can define $P = (p_{00}, p_{10}, p_{01}, p_{11})$ and $Q = (q_{00}, q_{10}, q_{01}, q_{11})$ where

$$\begin{aligned}
p_{00} &= \mathbb{P}(Y = 0, D = 0), & q_{00} &= \mathbb{P}(Y_0 = 0, Y_1 = 0),\\
p_{10} &= \mathbb{P}(Y = 1, D = 0), & q_{10} &= \mathbb{P}(Y_0 = 1, Y_1 = 0),\\
p_{01} &= \mathbb{P}(Y = 0, D = 1), & q_{01} &= \mathbb{P}(Y_0 = 0, Y_1 = 1),\\
p_{11} &= \mathbb{P}(Y = 1, D = 1), & q_{11} &= \mathbb{P}(Y_0 = 1, Y_1 = 1).
\end{aligned}$$

The correspondence $G : \mathcal{U} \to \mathcal{Y}$ for this model is defined by:

$$G(y_0, y_1) = \{(y, d) : y = y_1 d + (1 - d) y_0\}.$$

The capacity functional associated with the correspondence G^{-1} is given by:

$$\begin{aligned}
T(\{(0,0)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,0)\}) &&= p_{00} + p_{01}, &(6.1)\\
T(\{(0,1)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,1)\}) &&= p_{00} + p_{11}, &(6.2)\\
T(\{(1,0)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(1,0)\}) &&= p_{10} + p_{01}, &(6.3)\\
T(\{(1,1)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(1,1)\}) &&= p_{10} + p_{11}, &(6.4)\\
T(\{(0,0),(0,1)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,0),(0,1)\}) &&= p_{00} + p_{01} + p_{11}, &(6.5)\\
T(\{(0,0),(1,0)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,0),(1,0)\}) &&= p_{00} + p_{01} + p_{10}, &(6.6)\\
T(\{(0,0),(1,1)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,0),(1,1)\}) &&= p_{00} + p_{01} + p_{10} + p_{11} = 1, &(6.7)\\
T(\{(0,1),(1,0)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,1),(1,0)\}) &&= p_{00} + p_{01} + p_{10} + p_{11} = 1, &(6.8)\\
T(\{(0,1),(1,1)\}) &= \mathbb{P}(G(Y, D)^{-1} \cap \{(0,1),(1,1)\}) &&= p_{00} + p_{10} + p_{11}, &(6.9)
\end{aligned}$$

$$\begin{aligned}
T(\{(1,0),(1,1)\}) &= \mathbb{P}(G(Y,D)^{-1}\cap\{(1,0),(1,1)\}) &&= p_{01}+p_{10}+p_{11}, && (6.10)\\
T(\{(0,0),(0,1),(1,0)\}) &= \mathbb{P}(G(Y,D)^{-1}\cap\{(0,0),(0,1),(1,0)\}) &&= p_{00}+p_{01}+p_{10}+p_{11}=1, && (6.11)\\
T(\{(0,0),(0,1),(1,1)\}) &= \mathbb{P}(G(Y,D)^{-1}\cap\{(0,0),(0,1),(1,1)\}) &&= p_{00}+p_{01}+p_{10}+p_{11}=1, && (6.12)\\
T(\{(0,0),(1,0),(1,1)\}) &= \mathbb{P}(G(Y,D)^{-1}\cap\{(0,0),(1,0),(1,1)\}) &&= p_{00}+p_{01}+p_{10}+p_{11}=1, && (6.13)\\
T(\{(0,1),(1,0),(1,1)\}) &= \mathbb{P}(G(Y,D)^{-1}\cap\{(0,1),(1,0),(1,1)\}) &&= p_{00}+p_{01}+p_{10}+p_{11}=1, && (6.14)\\
T(\mathcal{U}) &= \mathbb{P}(G(Y,D)^{-1}\cap\mathcal{U}) &&= p_{00}+p_{01}+p_{10}+p_{11}=1. && (6.15)
\end{aligned}$$

Given the capacity functional of the random set $G^{-1}(Y,D)$, it is possible to characterize the set of all distributions $Q\in\mathcal{Q}^\dagger$ that are consistent with the observed distribution P. Denote the set of all such distributions as the set $\mathcal{Q}$. A result from random set theory call *Artstein's theorem* (Artstein (1983)) —stated formally in Appendix A in the online supplementary material— provides us with the necessary and sufficient conditions for the existence of a random variable $\widetilde{U}$ with distribution $Q\in\mathcal{Q}^\dagger$ that can rationalize the observed distribution P through the model correspondence G. In particular, the necessary and sufficient conditions in Artstein's theorem for $\widetilde{U}\sim Q$ to be able to rationalize the observed distribution P are given by:

$$Q(A)\le T(A) \qquad \forall A\in 2^{\mathcal{U}}$$

I.e. they are expressed as a dominance condition of the distribution Q over the capacity functional T; see the discussion in Beresteanu et al. (2012). Since $\mathcal{U}$ is a finite set, to verify that a given distribution Q can rationalize the observed distribution P requires the researcher to check if a finite number of linear inequality constraints are satisfied. This implies that we can write our collection $\mathcal{Q}$ —the collection of $Q\in\mathcal{Q}^\dagger$ that are consistent with the observed distribution P— as:

$$\mathcal{Q}=\{Q\in\mathcal{Q}^\dagger : Q(A)\le T(A) \text{ for all } A\in 2^{\mathcal{U}}\} \tag{7}$$

If the constraints defining $\mathcal{Q}^\dagger$ are linear, then $\mathcal{Q}$ is a polyhedron contained in the $K-$dimensional probability simplex; see figure 1. This idea is illustrated in the following example.

Example 1 (Cont'd)**.** Consider again the case where $Y\in\{0,1\}$ and $D\in\{0,1\}$, and recall the capacity functional given by (6.1)-(6.15). Given the correspondence G, we know from Artstein's theorem that Q is a joint distribution function that can rationalize the observed distribution P if the following inequalities are satisfied:

$$\begin{aligned}
q_{00} &\le p_{00}+p_{01}, && (8.1)\\
q_{01} &\le p_{00}+p_{11}, && (8.2)\\
q_{10} &\le p_{10}+p_{01}, && (8.3)\\
q_{11} &\le p_{10}+p_{11}, && (8.4)
\end{aligned}$$

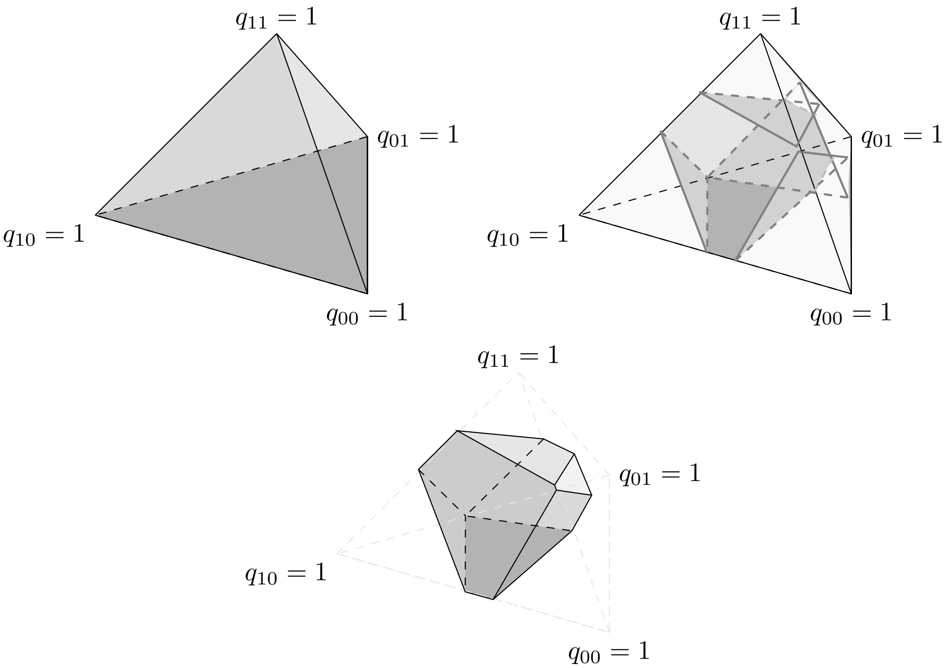


Figure 1: A representation of the constraints imposed by Artstein's theorem on the probability simplex in the case when $Y, D \in \{0, 1\}$.

$$q_{00} + q_{01} \leq p_{00} + p_{01} + p_{11}, \tag{8.5}$$
$$q_{00} + q_{10} \leq p_{00} + p_{01} + p_{10}, \tag{8.6}$$
$$q_{00} + q_{11} \leq 1, \tag{8.7}$$
$$q_{01} + q_{10} \leq 1, \tag{8.8}$$
$$q_{01} + q_{11} \leq p_{00} + p_{10} + p_{11}, \tag{8.9}$$
$$q_{10} + q_{11} \leq p_{01} + p_{10} + p_{11}, \tag{8.10}$$
$$q_{00} + q_{01} + q_{10} \leq 1, \tag{8.11}$$
$$q_{00} + q_{01} + q_{11} \leq 1, \tag{8.12}$$
$$q_{00} + q_{10} + q_{11} \leq 1, \tag{8.13}$$
$$q_{01} + q_{10} + q_{11} \leq 1, \tag{8.14}$$
$$q_{00} + q_{01} + q_{10} + q_{11} \leq 1. \tag{8.15}$$

The following theorem shows how the characterization of the collection $\mathcal{Q}$ can be used to bound continuous functionals of the joint distribution $f : \mathcal{Q} \to \mathbb{R}$.

Theorem 1. *Let $G : \mathcal{U} \to \mathcal{Y}$ be a correspondence and define*

$$\mathcal{Q} = \{Q \in \mathcal{Q}^\dagger : Q(A) \leq T(A) \text{ for all } A \in 2^{\mathcal{U}}\},$$

where $\mathcal{Q}^\dagger$ is a convex set of admissible distributions. If $\mathcal{Q}$ is nonempty, then for every continuous functional

$f : \mathcal{Q} \to \mathbb{R}$, the identified set for f is a nonempty interval $[f^\ell, f^u]$ where:

$$f^u = \sup_{Q\in\mathcal{Q}} f(Q), \qquad\qquad f^\ell = \inf_{Q\in\mathcal{Q}} f(Q). \tag{9}$$

The intuition is straightforward. The collection $\mathcal{Q}$ provides us *all* distributions Q on $\mathcal{U}$ that can rationalize the observed distribution P on $\mathcal{W}$. Thus, to bound a function of the joint distribution Q, we need only to search over the set $\mathcal{Q}$ for the distributions that minimize and maximize our function of interest. Compactness of $\mathcal{Q}$ and continuity of f then guarantees that $\arg\min f(Q) \in \mathcal{Q}$ and $\arg\max f(Q) \in \mathcal{Q}$, and that the identified set for f is an interval.

Although the theorem is stated for the case when G is a correspondence —so as to accommodate the POM— it applies equally to the case when G is a function (i.e. when the model is *complete*). In addition, although we have stated the theorem using the capacity functional defined in equation (4) on the observables, the theorem could have been written in an analogous manner using the capacity functional (5) defined on the unobservables.

Note that theorem 1 is of interest from a practical point of view since —although it is an identification result— it suggests a straightforward method of estimation (the consistency of which is proven in Appendix D of the online supplemental material). Indeed, since the restrictions that define $\mathcal{Q}$ are linear in many cases, there exists a wide range of functions f for which computing f^u and f^ℓ reduces to solving a linear programming problem. The efficiency with which linear programs can be solved in the presence of a large number of constraints makes them particularly useful when computing bounds in partial identification. Even in cases when f is not linear, an increasing f, or concave/convex f can also lead to optimization problems that can be solved efficiently.[9] Note that this variational representation of the problem has significant advantages over analytic characterizations. Indeed, analytic characterizations must be derived for each parameter separately to ensure that they exploit all the information under the set of model assumptions. Analytic characterizations may quickly become unreasonable to provide, and for many interesting parameters, analytic characterizations simply do not yet exist, even in very simple environments.[10] In addition, even when plausible analytic bounds are proposed for such parameters, proving that the analytic characterization is *sharp* can be challenging in many circumstances. In contrast, theorem 1 provides a variational representation that we know produces sharp bounds, since by construction the bounds must respect all of the restrictions implied by the data on the distribution of unobservables. By imposing constraints on $\mathcal{Q}^\dagger$, the results also extend easily to accommodate additional modeling assumptions, such as the monotone treatment response, and monotone instrumental variables assumptions (see the discussion in Appendix C of the online supplemental material).

Note that bounds on many different functionals $f(Q)$ can be computed *without modifying the set of constraints* in the program defined by (9); i.e. once the constraints for the model correspondence G have

[9] Even if f does not meet these criteria, nonlinear optimization problems subject to linear constraints can be solved very quickly by many software applications, including Matlab, when the gradient of f is provided to the solver.

[10] For example, in the presence of an instrument there currently exists no analytic sharp bounds for the parameter $Q(Y_d \in A|Y_{1-d} \in B)$, even when Y and D are binary.

been established, the researcher is able to compute sharp bounds for many objects by simply changing the objective function. This result is a remarkable improvement over analytic characterizations which would require deriving bounds for all examples of $f(Q)$ above, and then proving the sharpness of the derived bounds. As we will see in the application in section 4, this feature will allow the researcher to easily compute a variety of causal parameters to give a complete view of the effects of a program.

2.3 Identification with an Instrument

With some abuse of notation, we can redefine the model as $(G, \mathcal{Q}^\dagger)$, where $\mathcal{Q}^\dagger$ is the collection of all admissible distributions and G is the model correspondence $G : \mathcal{U} \times \mathcal{Z} \to \mathcal{W} \times \mathcal{Z}$. In our case, the model correspondence for the POM with an instrument is given by:

$$G(y_0, y_1, \ldots, y_{K-1}, z) = \left\{ (y, d, z) \in \mathcal{W} \times \mathcal{Z} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d = k\} \right\}, \tag{10}$$

for $(y_0, y_1, \ldots, y_{K-1}, z) \in \mathcal{U} \times \mathcal{Z}$. Note that —similar to the case without an instrument— this model is also *incomplete* in the sense that the process generating the (possibly endogenous) variable D has not been specified. Importantly, we do not impose any structure between D and Z, or D and any of the other unobservable variables.

Given the assumption that $Z \perp\!\!\!\perp U$, for any $Q \in \mathcal{Q}^\dagger$ we must have $Q(A|Z = z) = Q(A)$. For each $z \in \mathcal{Z}$, we can define the conditional capacity functional:

$$T(A|Z = z) = P(G^{-1}(Y, D) \cap A \neq \emptyset | Z = z),$$

for every $A \in 2^{\mathcal{U}}$. Let $\mathcal{Q}_z$ denote the set of distributions that are admissible and also satisfy Artstein's inequalities subject to the conditional capacity functional:

$$\mathcal{Q}_z = \{Q \in \mathcal{Q}^\dagger : Q(A) \leq T(A|Z = z) \text{ for all } A \in 2^{\mathcal{U}}\}.$$

Note that since the probability measure Q must respect Artstein's inequalities *for all values of* $z \in \mathcal{Z}$, the identified set $\mathcal{Q}$ in the presence of an instrument can be written:

$$\mathcal{Q} = \bigcap_{z \in \mathcal{Z}} \mathcal{Q}_z. \tag{11}$$

The concept of this set is illustrated in figure 2. The construction of the identified set in this way in the presence of an instrument is discussed in Beresteanu et al. (2012), and a proof of its validity is provided in their proposition 2.5. It is also discussed at length by Chesher and Rosen (2017), with an analogous result to that in Beresteanu et al. (2012) provided by their theorem 4.

In a practical sense, this identified set is constructed by listing Artstein's inequalities for every value of $z \in \mathcal{Z}$, and then finding the distributions Q that respect all inequalities. The following example provides a sense of the method:

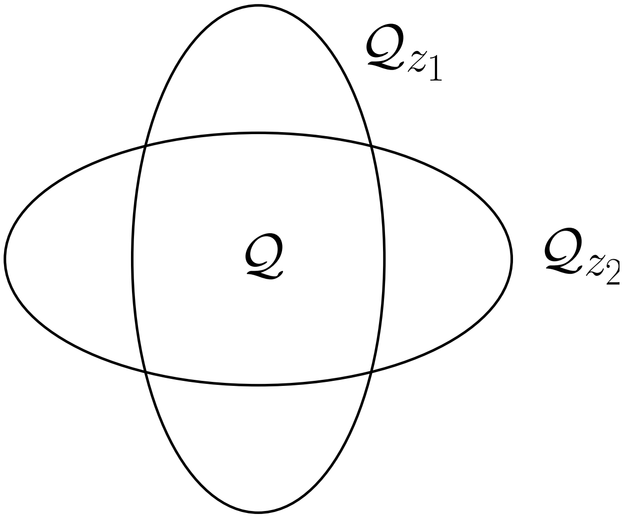


Figure 2: A representation of the identified set in the presence of an instrument $Z \in \{z_1, z_2\}$.

Example 2. Consider again the case of a binary outcome and binary treatment. Under this assumption, the distribution of (Y_0, Y_1) is "unaffected" by the presence of the instrument in the sense that $Q(Y_0 \in A, Y_1 \in B|Z = z) = Q(Y_0 \in A, Y_1 \in B)$. Since the outcome and treatment variable are binary we can still write the distribution of (Y_0, Y_1) as $Q = (q_{00}, q_{10}, q_{01}, q_{11})$; however, we must now define the conditional distribution $P(z) = (p_{00}(z), p_{10}(z), p_{01}(z), p_{11}(z))$, where $p_{ij}(z) = P(Y = i, D = j|Z = z)$. Then Artstein's inequalities can be written for each fixed $Z = z$ as:

$$\begin{aligned}
q_{00} &\leq p_{00}(z) + p_{01}(z),\\
q_{01} &\leq p_{00}(z) + p_{11}(z),\\
q_{10} &\leq p_{01}(z) + p_{10}(z),\\
q_{11} &\leq p_{10}(z) + p_{11}(z),\\
q_{00} + q_{01} &\leq p_{00}(z) + p_{01}(z) + p_{11}(z),\\
q_{00} + q_{10} &\leq p_{00}(z) + p_{01}(z) + p_{10}(z),\\
q_{00} + q_{11} &\leq 1,\\
q_{01} + q_{10} &\leq 1,\\
q_{01} + q_{11} &\leq p_{00}(z) + p_{10}(z) + p_{11}(z),\\
q_{10} + q_{11} &\leq p_{01}(z) + p_{10}(z) + p_{11}(z),\\
q_{00} + q_{01} + q_{10} &\leq 1,\\
q_{00} + q_{01} + q_{11} &\leq 1,\\
q_{00} + q_{10} + q_{11} &\leq 1,\\
q_{01} + q_{10} + q_{11} &\leq 1,\\
q_{00} + q_{01} + q_{10} + q_{11} &\leq 1.
\end{aligned}$$

Since Artstein's inequalities must hold for every $z \in \mathcal{Z}$, each inequality must hold for the infimum of the right hand side of the inequalities over the values of $z \in \mathcal{Z}$. Thus, rather than write the inequalities for each

z, we can equivalently write:

$$\begin{aligned}
q_{00} &\leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{01}(z)\}, && (12)\\
q_{01} &\leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{11}(z)\}, && (13)\\
q_{10} &\leq \inf_{z \in \mathcal{Z}} \{p_{01}(z) + p_{10}(z)\}, && (14)\\
q_{11} &\leq \inf_{z \in \mathcal{Z}} \{p_{10}(z) + p_{11}(z)\}, && (15)\\
q_{00} + q_{01} &\leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{01}(z) + p_{11}(z)\}, && (16)\\
q_{00} + q_{10} &\leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{01}(z) + p_{10}(z)\}, && (17)\\
q_{00} + q_{11} &\leq 1, && (18)\\
q_{01} + q_{10} &\leq 1, && (19)\\
q_{01} + q_{11} &\leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{10}(z) + p_{11}(z)\}, && (20)\\
q_{10} + q_{11} &\leq \inf_{z \in \mathcal{Z}} \{p_{01}(z) + p_{10}(z) + p_{11}(z)\}, && (21)\\
q_{00} + q_{01} + q_{10} &\leq 1, && (22)\\
q_{00} + q_{01} + q_{11} &\leq 1, && (23)\\
q_{00} + q_{10} + q_{11} &\leq 1, && (24)\\
q_{01} + q_{10} + q_{11} &\leq 1, && (25)\\
q_{00} + q_{01} + q_{10} + q_{11} &\leq 1.. && (26)
\end{aligned}$$

When we write Artstein's inequalities by taking the infimum over all $z \in \mathcal{Z}$ on the right hand side, we call this "intersecting" over the value of $z \in \mathcal{Z}$.

After enumerating the entire relevant set of Artstein's inequalities as in the above example, it is straightforward to see that the bounding procedure suggested by theorem 1 is then applicable to the case with an instrument where the identified set is as defined in equation (11); see Beresteanu et al. (2012) for additional discussion of this approach.

Note that there is no guarantee that the identified set defined by equation (11) is nonempty. In contrast, the identified set without an instrument is always nonempty if $\mathcal{Q}^{\dagger}$ is nonempty. By definition, emptiness of the identified set in (11) implies that there exists no random variable $U \in \mathcal{U}$ that can generate the observed distribution while respecting the condition $Z \perp\!\!\!\perp U$ and the restrictions on $\mathcal{Q}^{\dagger}$. Thus, when $\mathcal{Q}^{\dagger}$ is unrestricted, emptiness of the identified set provides evidence against the independence assumption. This intuition forms the basis for the test of independence proposed by Kédagni and Mourifie (2017).

2.4 Relation to Previous Work

In the treatment effect literature, Mourifie et al. (2015) provide sharp bounds on a variety of parameters — with and without an instrument— in the case of binary outcome and binary treatment. Theorem 1 provides a sharp characterization for parameters such as the distributional mobility parameter for which Mourifie et al. (2015) do not claim sharpness. In addition, Theorem 1 enables the bounds to be implemented easily for treatment effect models with arbitrary discrete-valued outcome and treatment rather than for just the binary outcome, binary treatment case focused on in Mourifie et al. (2015).

Theorem 1 is related to proposition 1 in Torgovitsky (2016), which provides an analogous result but for the class of complete econometric models.[11] Theorem 1 extends the result of Torgovitsky (2016) to incomplete econometric models using random set theory. These incomplete models include the POM (with and without instrument) that is of primary interest in this paper. Since incomplete models nest complete models, theorem 1 implies proposition 1 in Torgovitsky (2016) when G is a function rather than a correspondence.

The relationship between the approach based on Artstein's theorem and an alternative approach using the Aumann expectation in Beresteanu et al. (2011) is discussed at length in Beresteanu et al. (2012). However, while the authors consider Artstein's theorem for bounding probability distributions, they do not consider using Artstein's theorem to bound functionals of the joint distribution. In contrast to the Aumann-expectation approach, theorem 1 provides a bounding approach for discrete-valued outcomes, and requires solving only two optimization problems.

3 Efficient Computation

Although theorem 1 suggests a straighforward method of computing bounds on functionals of the joint distribution, it may not always be computationally feasible. To appreciate the computational burden implied by theorem 1, note that the identified set constructed via Artstein's inequalities is restricted by $2^{\min(|\mathcal{Y}|,|\mathcal{U}|)}-2$ constraints.[12] As noted by Beresteanu et al. (2012), when the support of the outcome variable is large, the number of inequalities that constrain the identified set can become prohibitive. For example, when $|\mathcal{Y}| = 20$, $|\mathcal{D}| = 2$, there are over a trillion constraints on the identified set (precisely, 1.1×10^{12}). This makes estimation computationally infeasible.

In this section we explore two additional methods of efficiently computing bounds on functionals of the joint distribution. The first method is related to the idea of a *core determining class* introduced by Galichon and Henry (2011). Informally, a core determining class is any collection $\mathcal{S}$ of subsets of $\mathcal{U}$ such that if Artstein's inequalities hold for every $A \in \mathcal{S}$ then they also hold for every $A \in 2^{\mathcal{U}}$. In this section I discuss the properties of the smallest known core determining class for the POM, and show the number of remaining

[11] Indeed, the model in Torgovitsky (2016) is complete. This is because of the way Torgovitsky (2016) solves the initial conditions problem in his analysis of state dependence. For example, in his leading case of a binary outcome, he models state dependence nonparametrically through the recursive model:

$$Y_{it} = Y_{it-1}U_{it}(1) + (1 - Y_{it-1})U_{it}(0) = U_{it}(Y_{it-1}) \qquad \forall t \geq 1, \tag{27}$$

where $Y_{it} \in \{0, 1\}$ is the outcome in period t, and $U_{it}(y)$ are the counterfactual states in period t if $Y_{it-1} = y$ is imposed exogenously. To solve the initial conditions problem, Torgovitsky (2016) imposes that $U_{i0} = Y_{i0}$. However, with U_{i0} known, it is straightforward to see from the recursive nature of the model given by (27) that a vector $U = (U_{i0}, U_{i1}(0), \ldots, U_{iT}(0), U_{i1}(1), \ldots, U_{iT}(1))$ uniquely determines that path of outcomes $\{Y_{it}\}_{t=0}^{T}$.

[12] We can take the minimum in the exponent since it is equivalent (from an identification perspective) to write Artstein's inequalities either on the observables using the capacity functional given by (5) or unobservables using the capacity functional given by (4).

non-redundant constraints on the identified set. The second method discussed in this section is based on reframing the bounding problem as an optimal transport problem. Such an approach was suggested in Galichon and Henry (2006, 2011) and is explored here for the specific case of the POM.

All results in this section are given for the case when an instrument is available. However, researchers should keep in mind that the conclusions in this section may change if more structure is added to the POM (such as specifying a selection mechanism), since the additional structure may change the model correspondence. In either case, the arguments of this section show how a researcher might determine the optimal computational approach.

3.1 The Core Determining Class Approach

The idea that Artstein's theorem may provide many redundant inequalities appears first in the concept of a *core determining class* introduced by Galichon and Henry (2011):

Definition 2. *Let $\mathbf{X} \subset \mathcal{X}$ be a random set on the finite space $\mathcal{X}$. Consider the capacity functional $T_{\mathbf{X}}(A)$ for any $A \in 2^{\mathcal{X}}$. A collection $\mathcal{S} \subseteq 2^{\mathcal{X}}$ is called core-determining for the random set $\mathbf{X}$ if for any random variable $X \in \mathcal{X}$:*

$$\mathbb{P}(X \in A) \leq T_{\mathbf{X}}(A),$$

holds for all $A \in \mathcal{S}$, then the same inequality holds for all $A \in 2^{\mathcal{U}}$.

Note that the core determining class is defined for each random set, and so will be different depending on the model. In our context, the random set of interest is the set $G^{-1}(Y, D)$. Thus a core determining class will be any collection $\mathcal{S}$ of sets $A \in 2^{\mathcal{U}}$ such that if $Q(A) \leq T(A)$ holds for all all $A \in \mathcal{S}$, then the same inequality holds for all $A \in 2^{\mathcal{U}}$. This definition is consistent with the definition presented in Galichon and Henry (2011) and Chesher and Rosen (2017). From this definition, any $A \in 2^{\mathcal{U}}$ with $A \notin \mathcal{S}$ imposes a redundant constraint on the characterization of the identified set.

It is important to note that as given in definition 2, the core determining class is not unique, and thus a core determining class may still generate redundant inequalities via Artstein's theorem. Thus, to eliminate the largest number of redundant constraints, we are interested in finding the smallest possible core determining class. In a recent paper, Luo and Wang (2016) present conditions that allow for the elimination of redundant constraints implied by Artstein's theorem and, to the best of our knowledge, they provide the smallest available core determining class. Luo and Wang (2016) call their core determining class the *exact core determining class.* Using the specific structure of the correspondence for the POM and the mathematical results of Luo and Wang (2016), we characterize both the precise *number* and *type* of sets in the exact core determining class for the POM.[13] In particular, we are able to show that the number of restrictions on the joint distribution implied by the exact core determining class is small relative to the number of restrictions

[13] Note that the exact core determining class of Luo and Wang (2016) does not depend on whether we define Artstein's inequalities on the observables or the unobservables; for a given problem, the size of their core determining class is fixed.

implied by Artstein's theorem (although recall the exact core determining class contains the same sharp information as Artstein's inequalities).

Results on the precise nature of sets in the exact core determining class in the POM are given in lemmas 1, 2 and 3, which have been moved to Appendix B in the online supplemental material for brevity. Appendix B also contains a larger discussion of the results of Luo and Wang (2016), and how they were used to derive the results in this paper. However, using lemmas 1, 2 and 3, it is possible to show the following result.[14]

Proposition 1. *In the POM with an instrument there are*

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

sets in the exact core determining class.

This proposition provides us with the number of sets in the exact core determining class for the POM, and will help us compare the exact core's computational tractability to other methods. The precise characterization of the non-redundant sets in the exact core determining class are provided in Appendix B, and the results ensure that the redundant inequalities can be efficiently identified and removed from the bounding problem.

However, it is also important to note that under the exact core determining class approach of Luo and Wang (2016), it is not possible to compute the non-redundant constraints and then "intersect" by taking the infimum over all values of $z \in \mathcal{Z}$, as in the inequalities (12)-(26). Intuitively, this is because the results of Luo and Wang (2016) are valid only when the observed distribution P is a proper probability measure, and the infimum of $P(\cdot|Z = z)$ over values of $z \in \mathcal{Z}$ is generally not a probability measure.[15] Because of this feature, we will see that there are situations where the core determining class approach can be less computationally efficient then Artstein's inequalities.

3.2 The Dual Approach

In this subsection we show how the duality result of Galichon and Henry (2006, 2011) between Artstein's inequalities and the existence of an *equilibrium selection mechanism* can be used to construct an efficient estimation method.[16] To this end, let $\mathcal{M}_{G|z}(P_z, Q)$ denote the set of Borel probability distributions, conditional

[14] The full version of proposition 1 is given in Appendix B.

[15] This is why each term in proposition 1 is multiplied by a factor of $|\mathcal{Z}|$, since under the core determining class approach we must enumerate the non-redundant inequalities for each $z \in \mathcal{Z}$ rather than intersecting over $z \in \mathcal{Z}$.

[16] We call this the "dual approach" in the spirit of Galichon and Henry (2006), who show that if $\mathcal{M}(P, Q)$ represents the set of Borel probability measures with marginals P and Q with support on $Graph(G) \equiv \{(u, w) : w \in G(u)\}$, then:

$$\sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi[-\mathbb{1}\{W \notin G(U)\}] = 0 \qquad \Longleftrightarrow \qquad \inf_{B \in 2^{\mathcal{U}}} [Q(B) - P(G^{-1}(W) \cap B \neq \emptyset)] = 0.$$

Indeed, the problems on the left and right can be shown to be dual optimal transport problems. It is easy to see that the problem on the right defines the set of all distributions Q that satisfy Artstein's inequalities. The dual problem on the left is the one discussed in this subsection.

on $Z = z$, with marginals $P_z \equiv P(Y, D|Z = z)$ and Q with support on $Graph(G)$ given by:

$$Graph(G) \equiv \{(u, w) : w \in G(u)\}.$$

Then we can define the following set:

$$\mathcal{Q}_z^* = \{Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}_{G|z}(P_z, Q)\}. \tag{28}$$

In other words, $\mathcal{Q}$ defines the set of all distributions Q such that there exists a joint distribution $\pi \in \mathcal{M}_{G|z}(P_z, Q)$ that can rationalize the observed distribution P_z through the correspondence G.[17] Define:

$$\mathcal{Q}^* \equiv \bigcap_{z \in \mathcal{Z}} \mathcal{Q}_z^*$$

This collection will be connected to the collection $\mathcal{Q}$ characterized by Artstein's inequalities through the following result:

Proposition 2. *Let $G : \mathcal{U} \to \mathcal{Y}$ be a correspondence, and define:*

$$\mathcal{Q}_z = \{Q \in \mathcal{Q}^\dagger : Q(A) \leq T(A|Z = z) \text{ for all } A \in 2^{\mathcal{U}}\}.$$

Furthermore, define:

$$\mathcal{Q}_z^* = \{Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}_{G|z}(P_z, Q)\}.$$

Then $\mathcal{Q}_z = \mathcal{Q}_z^*$, so that $\mathcal{Q} = \mathcal{Q}^*$.

The proof of this proposition follows directly after applying theorem 3 in Galichon and Henry (2011). The proposition shows that, from an identification perspective, verifying whether a given distribution Q satisfies Artstein's inequalities is equivalent to showing the existence of a joint distribution π that can rationalize the marginals P_z (for all $z \in \mathcal{Z}$) and Q. From a practical perspective, this duality result provides an alternative method of estimation by showing that:

$$f^u \equiv \sup_{Q \in \mathcal{Q}} f(Q) = \sup_{Q \in \mathcal{Q}^*} f(Q), \qquad f^\ell \equiv \inf_{Q \in \mathcal{Q}} f(Q) = \inf_{Q \in \mathcal{Q}^*} f(Q). \tag{29}$$

where $\mathcal{Q}^*$ is as defined in proposition (2). In fact, this method was used by Lafférs (2013a, 2015) to bound the average treatment effect. The following example shows how to implement this approach:

Example 1 (Cont'd)**.** Consider again the case where $Y \in \{0, 1\}$ and $D \in \{0, 1\}$, and suppose we have access to an instrument $Z \in \{0, 1\}$ satisfying $Z \perp\!\!\!\perp (Y_0, Y_1)$. Recall that $p_{ij}(z) = \mathbb{P}(Y = i, D = j|Z = z)$ and $q_{ij} = \mathbb{P}(Y_0 = i, Y_1 = j)$. In addition, recall the set of non-redundant inequalities implied by Artstein's theorem for this model:

[17] In the presence of the set-valued mapping G, this characterization is equivalent to establishing, for each candidate $Q \in \mathcal{Q}^\dagger$, the existence of an *equilibrium selection mechanism*.

$$
\begin{aligned}
q_{00} &\leq p_{00}(z) + p_{01}(z), && (30.1)\\
q_{01} &\leq p_{00}(z) + p_{11}(z), && (30.2)\\
q_{10} &\leq p_{10}(z) + p_{01}(z), && (30.3)\\
q_{11} &\leq p_{10}(z) + p_{11}(z), && (30.4)\\
q_{00} + q_{01} &\leq p_{00}(z) + p_{01}(z) + p_{11}(z), && (30.5)\\
q_{00} + q_{10} &\leq p_{00}(z) + p_{01}(z) + p_{10}(z), && (30.6)\\
q_{01} + q_{11} &\leq p_{00}(z) + p_{10}(z) + p_{11}(z), && (30.7)\\
q_{10} + q_{11} &\leq p_{01}(z) + p_{10}(z) + p_{11}(z), && (30.8)
\end{aligned}
$$

where each inequality must hold for all $z \in \{0, 1\}$. In combination with the constraint $q_{ij} \geq 0$, we have that these constraints provide a sharp characterization of the identified set $\mathcal{Q}$.

Now let $\pi(z)$ be a vector with typical entry $\pi_{ijk\ell}(z)$, and consider the constraints:

$$
\begin{aligned}
\pi_{00,00}(z) + \pi_{00,01}(z) &= p_{00}(z), && (31.1)\\
\pi_{01,00}(z) + \pi_{01,10}(z) &= p_{01}(z), && (31.2)\\
\pi_{10,10}(z) + \pi_{10,11}(z) &= p_{10}(z), && (31.3)\\
\pi_{11,01}(z) + \pi_{11,11}(z) &= p_{11}(z), && (31.4)
\end{aligned}
$$

where $\pi_{ij,k\ell}(z) \geq 0$. Note that for any vector $\pi(z)$ satisfying (31.1)-(31.4), we can recover the vector $\mathbf{q}$ via:

$$
\begin{aligned}
\pi_{00,00}(z) + \pi_{01,00}(z) &= q_{00}, && (32.1)\\
\pi_{00,01}(z) + \pi_{11,01}(z) &= q_{01}, && (32.2)\\
\pi_{01,10}(z) + \pi_{10,10}(z) &= q_{10}, && (32.3)\\
\pi_{10,11}(z) + \pi_{11,11}(z) &= q_{11}. && (32.4)
\end{aligned}
$$

Since we have assumed $Z \perp\!\!\!\perp (Y_0, Y_1)$, we must ensure that the vectors $\pi(0)$ and $\pi(1)$ can generate the same probability vector $\mathbf{q}$. Thus in addition to imposing constraints (31.1)-(31.4) for $z = 0$ and $z = 1$, we must also impose a "compatibility restriction" given by:

$$
\begin{aligned}
\pi_{00,00}(z) + \pi_{01,00}(z) &= \pi_{00,00}(z') + \pi_{01,00}(z'), && (33.1)\\
\pi_{00,01}(z) + \pi_{11,01}(z) &= \pi_{00,01}(z') + \pi_{11,01}(z'), && (33.2)\\
\pi_{01,10}(z) + \pi_{10,10}(z) &= \pi_{01,10}(z') + \pi_{10,10}(z'), && (33.3)\\
\pi_{10,11}(z) + \pi_{11,11}(z) &= \pi_{10,11}(z') + \pi_{11,11}(z'), && (33.4)
\end{aligned}
$$

where $z, z' \in \{0, 1\}$. By proposition 2, the set of probability vectors $\mathbf{q}$ satisfying Artstein's inequalities for all z is equal to the set of probability vectors $\mathbf{q}$ recovered through (32.1)-(32.4) from any probability vectors $\pi(z)$, for $z = 0, 1$, satisfying (31.1)-(31.4) and the compatibility restrictions (33.1)-(33.4).

Given an instrument Z, a simple calculation shows that for the dual result in the presence of an instrument there are $|\mathcal{Y}|^{|\mathcal{D}|} \cdot |\mathcal{D}| \cdot |\mathcal{Z}|$ parameters, and $|\mathcal{Y}| \cdot |\mathcal{D}| \cdot |\mathcal{Z}| + (|\mathcal{Z}| - 1) \cdot |\mathcal{Y}| \cdot |\mathcal{D}|$ constraints.[18]

3.3 Comparison

We compute the number of constraints and parameters under different environments to provide a comparison of each characterization (Artstein's inequalities, the exact core, and the dual approach). First we consider the case where $|\mathcal{D}| = |\mathcal{Z}| = 2$, and we vary the cardinality of the support $\mathcal{Y}$; the results for this case are displayed in table 1. Second, we consider the case where $|\mathcal{D}| = |\mathcal{Y}| = 2$, and we vary the cardinality of the support $\mathcal{Z}$; the results for this case are displayed in table 2.

		$|\mathcal{Y}|$=2	$|\mathcal{Y}|$=3	$|\mathcal{Y}|$=4	$|\mathcal{Y}|$=5	$|\mathcal{Y}|$=10	$|\mathcal{Y}|$=20
Artstein	Parameters	4	9	16	25	100	400
	Constraints (Obs.)	15	63	255	1,023	1.0×10^{6}	1.1×10^{12}
	Constraints (Unobs.)	15	511	65,535	3.4×10^{7}	1.3×10^{30}	2.6×10^{120}
Artstein (Exact Core)	Parameters	4	9	16	25	100	400
	Constraints	16	54	192	550	40,680	8.4×10^{7}
Dual Problem	Parameters	16	36	64	100	400	1,600
	Constraints	12	18	24	30	60	120

Table 1: Number of parameters and non-redundant constraints from Artstein's theorem, the smallest core, and the dual problem in the presence of an instrument (excluding non-negativity constraints) where $D, Z \in \{0, 1\}$.

		$|\mathcal{Z}|$=200	$|\mathcal{Z}|$=300	$|\mathcal{Z}|$=400	$|\mathcal{Z}|$=500	$|\mathcal{Z}|$=1000	$|\mathcal{Z}|$=2000
Artstein	Parameters	4	4	4	4	4	4
	Constraints (Obs.)	15	15	15	15	15	15
	Constraints (Unobs.)	15	15	15	15	15	15
Artstein (Exact Core)	Parameters	4	4	4	4	4	4
	Constraints	1600	2400	3200	4000	8000	16000
Dual Problem	Parameters	1600	2400	3200	4000	8000	16000
	Constraints	1596	2396	3196	3996	7996	15996

Table 2: Number of parameters and non-redundant constraints from Artstein's theorem, the smallest core, and the dual problem in the presence of an instrument (excluding non-negativity constraints) where $D, Y \in \{0, 1\}$.

Table 1 shows that when the support $\mathcal{Y}$ has large cardinality, the number of constraints implied by Artstein's theorem and the exact core can be prohibitively large. In contrast, the dual approach implies a much smaller number of constraints, but a larger number of parameters. The reduction in the number of constraints afforded by the dual approach is found to have a significant impact on computational time; indeed, unreported simulations show that the dual approach tends to be much faster when D and Z have small support and Y has large support.

However, when the support of the instrument Z is large and the support of Y and D are small the dual approach may no longer be optimal. Table 2 shows that when the support of the instrument Z is large the dual approach requires significantly more parameters than either Artstein's inequalities or the exact core approach. In addition, both the exact core and dual approach require a similar number of constraints.

[18] Here I do not count non-negativity constraints.

However, the number of parameters and the number of constraints implied by Artstein's theorem remains constant: this is because —unlike the other approaches— Artstein's inequalities can be "intersected" over values of $z \in \mathcal{Z}$, as shown in the inequalities (12)-(26). This property unique to Artstein's inequalities make them especially computationally tractable when the cardinality of Z is large.

Also note that neither table 1 or 2 seem to support the use of the exact core approach, which is either dominated by the dual approach (in table 1) or by Artstein's inequalities (in table 2). When trying other combinations of $|\mathcal{D}|$, $|\mathcal{Z}|$ and $|\mathcal{Y}|$ we were unable to find environments where the exact core approach was clearly dominant, although there were many situations when its computational time was comparable to either the dual approach or Artstein's inequalities.

Overall, using a program that chooses the optimal approach for the problem at hand (either Artstein's inequalities, the exact core approach, or the dual approach) is found to alleviate a significant amount of the computational burden associated with the optimization problems in theorem 1, making the approach in this paper tractable to run on a standard laptop computer for many bounding problems.

4 Application

We apply the results in this paper to the well-known Tennessee STAR experiment analyzed in Krueger (1999) and Krueger and Whitmore (2001). Beginning in 1985, the Tennessee STAR experiment was a longitudinal study looking to analyze the impact of class size on the academic performance of students. The study saw students and teachers randomized within schools into classrooms of varying sizes; small classrooms had 13-17 students, and regular classrooms had between 22-25 students. The regular classrooms were divided between regular classrooms with and without a teacher aide. The objective of the study was to evaluate the impact of reduced class sizes on students' performance on standardized tests for reading and math. The initial random assignment of students to classrooms was done within schools at the kindergarten level. Students were then expected to respect their initial assignment for four years (i.e. until the end of grade 3). A detailed background of the study is provided in Finn et al. (2007).

As discussed in Krueger (1999), although the initial assignment of students to classrooms was random, the study was affected by a number of experimental issues. First, although nearly all students respected the initial assignment, approximately 10 percent of students switched between small and regular classrooms between each grade. Second, there is evidence of a significant amount of attrition in the sample; Krueger (1999) reports that half of the students present in kindergarten were missing in at least one subsequent year. Third, due to the (possibly non-random) attrition from the study and the natural movement of families in and out of areas that included a participating school, the actual range of class sizes differed from the initial experimental targets; for example, the true range of class sizes for "small" class sizes was 11-20 students, and the true range for "regular" class sizes was 15-30 students. Fourth, for children entering a school after kindergarten, the assignment of children to small or regular classrooms depended on the slots available in each classroom. As a result, the randomization for newly entering students was not perfectly balanced across classroom sizes. Finally, children assigned to regular classrooms were re-randomized each year into regular classes with a teacher aide and regular classes without a teacher aide. The result is that children initially assigned to small classrooms in kindergarten were more likely to stay with the same cohort of peers up to grade three. If the stability in the composition of a child's peers has an effect on academic performance, this effect may contribute to any differences between test scores of children in small versus regular classes.

We use the methods in this paper to provide various measures of the causal effects of the program on student performance. The outcome of interest will be the student's average class percentile ranking on reading and math exams administered in grade 3 and grade 8. Specifically, for grade 3 the outcome is the average percentile ranking on the math and reading sections of the Stanford Achievement Test (SAT), and for grade 8 the outcome is the average percentile ranking on the math and reading sections of the Comprehensive Test of Basic Skills (CTBS). The grade 8 outcome is included to evaluate the long-term impact of the program. The treatment indicator D is equal to 1 if a child has been in a classroom with ≤ 17 students for every grade before grade 4. Note that the *actual* class size —not the label of the class as small or regular— is used to construct the treatment variable. Also note that, due to the possibly non-random switching or assignment to small class sizes in the grades above kindergarten, the treatment variable may be correlated with potential outcomes.

The sample is restricted to those who participated in the STAR program in kindergarten, and for whom data on grade 3 and grade 8 test scores were available. The final sample size was $n = 2357$ students. Summary statistics for the selected sample are displayed in table 3. The table shows sample means and standard deviations broken down by treatment/control groups (i.e. $D = 1$ and $D = 0$) and groups based on the random assignment to small and regular classrooms (i.e. $Z = 1$ and $Z = 0$). The table displays information on the sample characteristics, and on the outcomes by subgroup. The reported outcomes are the *average student percentile ranks* (across all participating schools) for the reading and math test scores in grade 3 and grade 8. Note by construction the average percentile rank across the full sample will be exactly 50/100, so that the values in the table can be interpreted relative to this number. Notice in table 3 that sample characteristics are well-balanced across $Z = 1$ and $Z = 0$, which provides some evidence that the randomization was successful.[19] However, as we can see from the table, there was significant noncompliance to randomization. On average, percentile ranks appear to be higher both for students who were assigned to small classes, and for students who actually attended small classes from kindergarten to grade 3. Notice that across all tests and grades, average percentile ranks are higher for students who attended small classes from kindergarten to Grade 3 ($D = 1$) versus those students who were assigned to small classes in kindergarten ($Z = 1$); this provides some heuristic evidence of non-random selection between the small and regular classrooms.

For the purpose of the application, percentile ranks for math and reading scores were averaged together to

[19] Note that in the Tennessee STAR Experiment randomization occurred at the school level.

Table 3: Summary Statistics for the Selected Sample

		$D = 1^{\dagger}$	$D = 0^{\dagger}$	$Z = 1^{\dagger\dagger}$	$Z = 0^{\dagger\dagger}$
Sample Characteristics	Poor	0.34	0.35	0.35	0.34
		(0.47)	(0.48)	(0.48)	(0.48)
	Female	0.53	0.56	0.54	0.55
		(0.50)	(0.50)	(0.50)	(0.50)
	Black	0.21	0.25	0.24	0.24
		(0.41)	(0.43)	(0.43)	(0.43)
Average Reading Percentile*	Grade 3	55.72	48.38	54.66	48.00
		(28.33)	(28.82)	(28.77)	(28.68)
	Grade 8	52.66	49.23	51.96	49.18
		(28.77)	(28.86)	(28.83)	(28.86)
Average Reading Percentile*	Grade 3	54.73	48.67	53.56	48.48
		(28.59)	(28.81)	(28.75)	(28.79)
	Grade 8	53.55	49.00	51.95	49.18
		(28.89)	(28.79)	(28.78)	(28.88)
	Observations	527	1830	716	1641

†: $D = 1$ if student attends a small class from kindergarten to Grade 3
††: $Z = 1$ if student is randomly assigned to a small class in kindergarten
* : Note that by definition the average percentile rank for the full sample is 50 for both the reading and math test scores.

create a single outcome variable (as in Krueger (1999)). To reduce the computational burden for non-linear parameters —like the correlation and standard deviation— the percentiles were first discretized using the K-means algorithm. A variety of bin numbers were used (i.e. $25, 30$, and 35 bins) to show robustness of the results to the discretization. Also, given evidence from Krueger (1999) that the effect of a teacher aide was minimal, we consider the regular classes and the regular classes with an aide as simply 'regular classes,' and we evaluate the effect of treatment against this combined group. Finally, to obtain informative bounds, we impose a relaxed version of the monotone treatment response (MTR) assumption. Specifically, we impose that $\mathbb{P}(Y_1 > Y_0) \geq 0.95$. This implies that we consider only DGPs under which students strictly benefit from small classes sizes with 95% probability. This corresponds with our prior belief that smaller class sizes will be beneficial to most students, consistent with the evidence in Krueger (1999) and Krueger and Whitmore (2001) on the impact of the Tennessee STAR program.

To illustrate the flexibility of the method we provide bounds for the following parameters. Note that many of the parameters are constructed relative to the median percentile rank in the sample; this parameter is denoted Median(Y), where the median is taken *after* the K-means algorithm is applied to the outcome variable (and as such Median(Y) may differ slightly from 50/100).

(i) $\mathbb{P}(Y_0 \leq Median(Y), Y_1 > Median(Y))$*: The joint probability of having a percentile rank in the regular classroom (*Y_0*) lower than the median rank of the observed outcome* Y *and a percentile rank in the small classroom (*Y_1*) greater than the median rank of the observed outcome* Y*. Note that Median(Y) is only the median of the observed outcome, and is not necessarily the median of the counterfactual outcomes* Y_0 *and* Y_1*. As such, this parameter can provide a measure of symmetry in the joint distribution of unobserved counterfactual outcomes.*

(ii) $\mathbb{E}[Y_1 - Y_0]$*: The average treatment effect, which measures the average gain in rank a result of attending a small versus regular classroom.*

(iii) $\mathbb{P}(Y_1 > Y_0)$*: The voting criterion, which measures the proportion of students whose rank strictly improves from attending smaller classrooms. An important criteria for any policy, it will be important to ensure a significant number individuals benefit from the treatment. However, note that this parameter does not consider the identity of the individuals benefiting from treatment, and so does not capture distributional mobility or equality concerns.*

(iv) $\mathbb{P}(Y_1 > Median(Y)|Y_0 \leq Median(Y))$*: The conditional probability of being above the median rank in the small classroom given the individual is below the median rank in the regular classroom. From a policy perspective, this gives a measure of distributional mobility, as it shows the proportion of below-median students who can transition to becoming above-median students as a result of being in smaller classroom.*

(v) $\mathbb{P}(Y_0 \leq Median(Y))$*: The proportion of people who would have a below median rank in the regular classroom. Note again that Median(Y) is only the median percentile rank in the observed outcome, and not necessarily the median of the unobserved outcome* Y_0*. Therefore, if* $\mathbb{P}(Y_0 \leq Median(Y)) > 0.5$*, then we know that the median of the unobserved rank* Y_0 *is less than the median of* Y*, and similarly if* $\mathbb{P}(Y_0 \leq Median(Y)) < 0.5$ *then we know the unobserved rank* Y_0 *is more than the median of* Y*. Using this method we could also recover information on other quantiles of the distribution of* Y_0*.*

(vi) $\mathbb{P}(Y_1 > Median(Y))$*: The proportion of people who would have an above-median rank in the small classroom. Note again that Median(Y) is only the median percentile rank in the observed outcome, and not necessarily the median of the unobserved outcome* Y_1*. Therefore, if* $\mathbb{P}(Y_1 > Median(Y)) > 0.5$*, then we know that the median of the unobserved rank* Y_0 *is more than the median of* Y*, and similarly if* $\mathbb{P}(Y_1 > Median(Y)) < 0.5$ *then we know the unobserved rank* Y_1 *is less than the median of* Y*. Using this method we could also recover information on other quantiles of the distribution of* Y_1*.*

(vii) $Corr(Y_0, Y_1)$*: The correlation between student ranks in regular versus small classrooms. This provides a direct measure of (linear) dependence of students' performance across treatment states. A positive correlation indicates that students with low ranks in regular class rooms are also likely to have low ranks in small classrooms. A positive correlation across treatment states will temper the observed effect of any policy designed to improve student outcomes.*

(viii) $\sqrt{Var(Y_1 - Y_0)}$*: The standard deviation of treatment effects, which is the standard deviation of the distribution of gains in rank as a result of attending a small versus regular classroom. This parameter provides a measure of the heterogeneity of treatment effects. A value close to zero is indicative of homogeneous treatment effects, whereas large positive values are indicate of highly heterogeneous treatment effects.*

Table 4: Bounds on School Achievement from the Tennessee STAR Experiment Assuming $\mathbb{P}(Y_1 > Y_0) \geq 0.95$

		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	0.07	0.25	0.03	0.23
	Bins=30	0.08	0.26	0.04	0.22
	Bins=35	0.09	0.26	0.04	0.23
$\mathbb{E}[Y_1 - Y_0]$:	Bins=25	4.80	19.27	1.42	17.03
	Bins=30	4.24	19.13	0.88	16.32
	Bins=35	3.99	18.32	0.95	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
$\mathbb{P}(Y_1 > Median(Y) \mid Y_0 \leq Median(Y))$:†	Bins=25	0.14	1.00	0.05	1.00
	Bins=30	0.15	1.00	0.07	0.99
	Bins=35	0.16	0.97	0.08	0.95
$\mathbb{P}(Y_0 \leq Median(Y))$:†	Bins=25	0.58	0.58	0.59	0.59
	Bins=30	0.55	0.55	0.52	0.52
	Bins=35	0.53	0.53	0.51	0.51
$\mathbb{P}(Y_1 > Median(Y))$:†	Bins=25	0.46	0.62	0.41	0.59
	Bins=30	0.50	0.66	0.49	0.65
	Bins=35	0.53	0.69	0.51	0.67
$Corr(Y_0, Y_1)$:	Bins=25	0.04	0.50	0.08	0.49
	Bins=30	0.04	0.50	0.07	0.50
	Bins=35	0.02	0.50	0.09	0.50
$\sqrt{Var(Y_1 - Y_0)}$:	Bins=25	2.38	28.11	2.94	27.03
	Bins=30	2.44	28.46	1.05	26.87
	Bins=35	2.07	28.12	1.13	27.52

$\dagger$: 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 $\mathbb{P}(Y_1 > Y_0) \geq 0.95$.

The bounds on the parameters above are estimated in Matlab using the Gurobi plug-in for the linear programs, and the KNITRO plug-in for the non-linear programs. The results of the analysis are displayed in table 4. First note that table 4 shows that the results are insensitive to the number of bins used in the discretization. Next, note that under the assumption of instrument independence and the MTR assumption we are able to obtain informative bounds on interesting parameters.

For the grade 3 outcomes in table 4, the joint probability $\mathbb{P}(Y_0 \leq Median(Y), Y_1 > Median(Y))$ is in the range $0.09 - 0.26$, meaning between 9 and 26 percent of the population have an unfavorable (below median) outcome in the untreated state, and a favorable (above median) outcome in the treated state. For the grade 8 outcomes the result is similar, with values in the range $0.04 - 0.23$. An extended analysis for this parameter is given in Appendix E in the online supplemental material.

For the average treatment effect, both the grade 3 and grade 8 bounds are informative, with ranges of $3.99 - 18.32$ percentile points and $0.95 - 16.27$ percentile points. These ranges indicate substantial benefits from attending small class sizes, and are consistent with the results of Krueger (1999) and Krueger and

Whitmore (2001).[20]

The voting criterion $\mathbb{P}(Y_1 > Y_0)$ is the only parameter estimated without imposing the MTR assumption.[21] For both the grade 3 and grade 8 results we find that the bounds on the voting criterion are generally large and uninformative in this application. Indeed, for the grade 3 outcomes we find that the proportion of students who strictly benefit from the program is in the range $0.11 - 0.98$. For the grade 8 outcomes it is in the range $0.05 - 0.98$.

Bounds on the conditional probability of transitioning to an average percentile rank above the median as a result of the program are also found to be wide and uninformative. For the grade 3 outcomes the bounds on $\mathbb{P}(Y_1 > Median(Y)|Y_0 \leq Median(Y))$ range from $0.16 - 0.97$ and for the grade 8 outcomes it ranges from $0.08 - 0.95$. An extended analysis for this parameter is also provided in Appendix E in the online supplemental material.

Bounds on the correlation coefficient are found to marginally informative, ranging from $0.02 - 0.5$ for the grade 3 outcomes and $0.09 - 0.50$ for the grade 8 outcomes. These positive and informative bounds are consistent with the intuition that the students who achieved a high percentile rank in small class sizes were also likely to have achieved a high percentile rank in regular class sizes.[22]

Finally, we consider bounds on the standard deviation of treatment effects. For the grade 3 outcomes we find a range of $2.07 - 28.12$ percentage points, and for the grade 8 outcomes we find a range of $1.13 - 27.52$ percentage points. This indicates that the data is consistent with a large range of variation of treatment effects, including values that are consistent with a significant amount of heterogeneity in the impact of the program.

For researchers interested in the sensitivity of the bounds to the MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$, Appendix E in the online supplemental material contains bounding results when this assumption is relaxed to $\mathbb{P}(Y_1 > Y_0) \geq 0.5$. As expected, the bounds on some parameters become less informative. Whether the bounds are informative in a particular application —and under which assumptions the bounds are informative— depends on the empirical context, and not on the method proposed in this paper (which will always deliver sharp bounds). More informative bounds can always be obtained by imposing additional assumptions, or additional restrictions on the selection mechanism, both of which can be accomplished under minor modifications of the presented method.

Overall the results are consistent with previous studies on the effects of the Tennessee STAR program, although they suggest that the conclusions on the effect of the program may be sensitive to the maintained assumptions. The application shows how the method in this paper can be used to identify bounds on

[20] In particular, the two-stage least squares estimates in Krueger (1999) indicate a reduction in class size of 10 students is associated with a 7 to 9 point increase in a student's average percentile ranking. Furthermore, Krueger and Whitmore (2001) find positive effects on middle school test scores, especially for students qualifying for the free lunch program in elementary school.

[21] This is because the MTR assumption *directly* restricts the voting criterion parameter, whereas it only indirectly restricts the other parameters.

[22] However, sensitivity analysis in Appendix E shows identification power for this parameter is likely coming from our MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$.

causal parameters —specifically parameters that depend on the joint distribution— that might be used as a robustness check in an analysis by demonstrating the (lack of) sensitivity of identification to the maintained assumptions.

5 Conclusion

This paper presents results on the identification and estimation of bounds on continuous functionals of the joint distribution of potential outcomes. For many interesting functionals the bounding problem is either a convex or linear program. The results were achieved by using the characterization of the identified set via Artstein's theorem from random set theory. In addition, alternative characterizations of the optimization problems were discussed that allow for efficient computation. The results extend easily to accommodate additional modeling assumptions, such as the monotone treatment response, and monotone instrumental variables assumptions (see the discussion in Appendix C). Finally, we show an application of the results to the Tennessee STAR experimental data.

Everything has been done for the case when the selection mechanism is left completely unspecified; showing examples of how the method can be used where the researcher imposes more structure on the selection mechanism is a logical next step. Imposing enough structure on the selection mechanism to complete the econometric model should yield substantial computational benefits and identification power. In this respect, the results of this paper pave the way for the development of future computational approaches to the partial identification of treatment effect parameters. The method may also find use in other fields. In particular, slight modification of the results in this paper make them applicable to a general class of incomplete econometric models. Researchers in other areas (e.g. empirical games, auctions) who are avid users of partial identification may also find these results interesting and useful. Applications of the results to other areas has not been explored, and remains as a topic of future research.

Supplementary Material

Included in the supplementary material is Appendix A on mathematical preliminaries of random set theory; Appendix B with a deep discussion of the core determining class approach; Appendix C with an example of implementing the bounding approach as a linear program, and a brief discussion of the monotone treatment response and monotone instrumental variables assumptions; Appendix D discussing consistency of the bounding method and suggestions for inference; Appendix E for additional results for the application; and Appendix F for the proofs of the main results.

References

Andrews, D. W. and Guggenberger, P. (2009). Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. *Econometric Theory*, 25(03):669–709.

Andrews, D. W. and Shi, X. (2014). Nonparametric inference based on conditional moment inequalities. *Journal of Econometrics*, 179(1):31–45.

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

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

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc.

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

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

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.

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.

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

Chernozhukov, V., Newey, W. K., and Santos, A. (2015). Constrained conditional moment restriction models. *arXiv preprint arXiv:1509.06311*.

Chesher, A. and Rosen, A. (2012). Simultaneous equations models for discrete outcomes: coherence, completeness, and identification.

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

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

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

Fan, Y., Guerre, E., and Zhu, D. (2017). Partial identification of functionals of the joint distribution of potential outcomes. *Journal of Econometrics*, 197(1):42–59.

Finn, J., Boyd-Zaharias, J., Fish, R., and Gerber, S. (2007). Project star and beyond: Database users guide. lebanon, tn: Heros.

Galichon, A. and Henry, M. (2006). Inference in incomplete models.

Galichon, A. and Henry, M. (2009). A test of non-identifying restrictions and confidence regions for partially identified parameters. *Journal of Econometrics*, 152(2):186–196.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, page rdr008.

Ginther, D. K. (2000). Alternative estimates of the effect of schooling on earnings. *Review of Economics and Statistics*, 82(1):103–116.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874.

Honoré, B. E. and Lleras-Muney, A. (2006). Bounds in competing risks models and the war on cancer. *Econometrica*, 74(6):1675–1698.

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

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

Kaido, H., Molinari, F., and Stoye, J. (2016). Confidence intervals for projections of partially identified parameters. *arXiv preprint arXiv:1601.00934*.

Kaido, H., Molinari, F., and Stoye, J. (2017a). Confidence intervals for projections of partially identified parameters. *arXiv preprint arXiv:1601.00934*.

Kaido, H., Molinari, F., Stoye, J., and Thirkettle, M. (2017b). Calibrated projection in matlab: Users' manual. *arXiv preprint arXiv:1710.09707*.

Kaido, H. and White, H. (2014). A two-stage procedure for partially identified models. *Journal of Econometrics*, 182(1):5–13.

Kédagni, D. and Mourifie, I. (2017). Generalized instrumental inequalities: Testing iv independence assumption.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28.

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

Laffërs, L. (2013b). Identification in models with discrete variables.

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

Manski, C. F. (2009). *Identification for prediction and decision.* Harvard University Press.

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.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Mourifie, I., Henry, M., and Meango, R. (2015). Sharp bounds for the roy model. Working Paper.

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.

ONLINE SUPPLEMENTARY MATERIAL

Appendix A 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 $\mathcal{X}$.[23] Let $\mathscr{B}(\mathcal{K})$ 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 $\mathbf{X} : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{K}, \mathscr{B}(\mathcal{K}))$.

Definition 3 (Random Closed Set (Molchanov (2005), pg. 1))**.** *The map* $\mathbf{X} : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{K}, \mathscr{B}(\mathcal{K}))$ *is called a random closed set if, for every compact set* A *in* $\mathscr{X}$*:*

$$\{\omega : \mathbf{X} \cap A \neq \emptyset\} \in \mathcal{F}$$

Definition 4 (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:

$$\{G^{-1}(Y, D) \cap A_1 \neq \emptyset\} \cap \{G^{-1}(Y, D) \cap A_2 \neq \emptyset\} \neq \emptyset,$$

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 5 (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:

[23] Note that since we consider a bounded subset $\mathscr{X} \subset \mathbb{R}^d$, all closed sets on $\mathscr{X}$ will be compact.

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:*

$$\mu(X \in A) \leq \nu(\mathbf{X} \cap A \neq \emptyset) \qquad \forall A \in \mathcal{K}(\mathcal{R}^d) \tag{34}$$

Appendix B 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 6 (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 will 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 7 (Luo and Wang (2016))**.** *Let S_u, S_y and S_y^{-1} be the collections of sets with the following properties:*

(a) S_u is the collection of all nonempty subsets $A \in 2^{\mathcal{U}}$, $A \neq \mathcal{U}$, such that

(i) A is self-connected.[24]

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

(b) S_w is the collection of all nonempty 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) S_w^{-1} is the collection of $A \subset \mathcal{U}$ and $A \neq \mathcal{U}$ such that there exists $B \subset S_w$ 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 theorem in Luo and Wang (2016) follows.

[24] 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 $G(A_1) \cap G(A_2) \neq \emptyset$.

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.[25] 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 $\mathcal{G} = (\mathcal{Y}, \mathcal{U}, G)$. Some appealing properties of the general bipartite graph $\mathcal{G}$ defined by the POM include:

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

(ii) Part $\mathcal{Y}$ 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 $y_1 \neq y_2$, we have $G^{-1}(y_1) \neq G^{-1}(y_2)$.

(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 $\mathcal{G}_k = (\mathcal{Y}_k, \mathcal{U}, G)$, where

[25] 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) $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ *for all* $i \neq j$

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

(iii) $G^{-1}(v) \cap G^{-1}(v') = \emptyset$ *for any pair* $(v, v') \in \mathcal{Y}_k$

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

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

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

(ii) $\mathcal{Y}_k \not\subseteq B$ *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 2 provide a complete characterization of the type of sets in $\mathcal{S}_u$ and $\mathcal{S}_w$, respectively, for the POM. In addition, lemma 3 gives conditions on when any $A \in \mathcal{S}_u$ is also in $\mathcal{S}_w^{-1}$. This information proves to be useful when constructing efficient algorithms to compute the exact core determining class for the POM. Indeed, the computational gains from knowing the structure of each of these collections is found to be large. These lemmas can then be used to prove proposition 1 in the main text. The full version of this proposition is provided below.

Proposition 1 (Full Version)**.** *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.

Appendix C 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 that 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:

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

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

$$\begin{aligned} q_{00} &= \pi_{00,00} + \pi_{00,01} \\ q_{01} &= \pi_{01,00} + \pi_{01,11} \\ q_{10} &= \pi_{10,01} + \pi_{10,10} \\ q_{11} &= \pi_{11,10} + \pi_{11,11} \end{aligned}$$

Then the optimization problem is:

$$\max_{\pi} \frac{\pi_{01,00} + \pi_{01,11}}{\pi_{00,00} + \pi_{00,01} + \pi_{01,00} + \pi_{01,11}} \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \pi = \mathbf{p} \\ \mathbf{0} \preccurlyeq \pi \preccurlyeq \mathbf{1} \end{cases} \tag{35}$$

To write this as a linear programming problem, define

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

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

Then the problem above can be re-written

$$\max_{\tilde{\pi},r} c' \cdot \tilde{\pi} \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \tilde{\pi} - \mathbf{p} \cdot r = 0 \\ d_1 \cdot \tilde{\pi} = 1 \\ d_2 \cdot \tilde{\pi} - r = 0 \\ \mathbf{0} \preccurlyeq \tilde{\pi} \preccurlyeq \mathbf{1} \\ r \geq 1 \end{cases} \tag{36}$$

This can be seen by replacing the objective function in (35) with the equivalent objective function in (36), by multiplying both sides of the constraint $\mathbf{A}_\pi \cdot \pi = \mathbf{p}$ in (35) 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

$$\max_{\tilde{\mathbf{q}}_r} c_r' \cdot \tilde{\mathbf{q}}_r \qquad s.t. \qquad \begin{cases} \mathbf{A}_r \cdot \tilde{\mathbf{q}}_r = \mathbf{a}_r \\ b_l \preccurlyeq \tilde{\mathbf{q}}_r \preccurlyeq b_u \end{cases} \tag{37}$$

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

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

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

The problem (37) 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.

C.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 framework. 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 8 (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 9 (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 (2013a, 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 proposition 3. 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.

Appendix D Consistency and Inference

So far we have assumed that the researcher has perfect knowledge of the probability distribution P. We now show that when we replace the probability measure P with it's sample analog $\hat{P}_n$ given by

$$\hat{P}_{n,z} \equiv \hat{P}_n((Y,D) \in A|Z=z) = \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{(Y_i,D_i) \in A, Z_i = z\}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{(Z_i = z\}},$$

it is possible to estimate the identified set consistently for any continuous functional f using the optimization problems presented in theorem 1 and, by their numerical equivalence, the alternative methods presented in the previous section.[26] To this end, define the set

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

or equivalently (by corollary 2):

$$\hat{\mathcal{Q}}_n \equiv \bigcap_{z\in\mathcal{Z}} \{Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}_z(\hat{P}_{n,z}, Q)\}.$$

I.e. $\hat{\mathcal{Q}}_n$ is the empirical analog of the set $\mathcal{Q}$, where each probability measure P_z has been replaced by $\hat{P}_{n,z}$. Consistency in the estimation of sets is usually defined in terms of the Hausdorff metric $d_H(\cdot,\cdot)$ defined for any two sets A and B as:

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

Here we are interested in establishing consistency with respect to the Hausdorff metric of the set

$$\hat{\Theta}_n^f = [\hat{f}_n^\ell, \hat{f}_n^u] \qquad \text{with} \qquad \hat{f}_n^\ell = \sup_{Q\in\hat{\mathcal{Q}}_n} f(Q), \qquad \hat{f}_n^u = \inf_{Q\in\hat{\mathcal{Q}}_n} f(Q) \tag{38}$$

for the set

$$\Theta^f = [f^\ell, f^u] \qquad \text{with} \qquad f^u = \sup_{Q\in\mathcal{Q}} f(Q), \qquad f^\ell = \inf_{Q\in\mathcal{Q}} f(Q). \tag{39}$$

Consistency is given in the following proposition:

Proposition 3. *Let $G : \mathcal{U} \to \mathcal{Y}$ be a correspondence, let $f : \mathcal{Q} \to \mathbb{R}$ be any continuous functional, and let $\hat{\Theta}_n^f$ and Θ^f be as defined in (38) and (39) respectively. Suppose that (a) $\mathcal{Q}^\dagger$ is closed and convex, (b) $\{W_i\}_{i=1}^n$ is an i.i.d. sample with $W_i = (Y_i, D_i) \in \mathcal{Y}$, (c) $\Theta^f \neq \emptyset$. Then $d_H(\hat{\Theta}_n^f, \Theta^f) \xrightarrow{p} 0$.*

Since f is a continuous functional, consistency follows from the continuous mapping theorem if we can show that $d_H(\hat{\mathcal{Q}}_n, \mathcal{Q}) \xrightarrow{p} 0$. To begin the proof, we first formulate the set of distributions in $\hat{\mathcal{Q}}_n$ as the set

[26] All results are presented with an instrument, since in the absence of an instrument the results will hold conditional on $Z =$ a constant.

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 criterion function fits into the framework of Shi and Shum (2015), and by verifying the conditions required for consistency presented in their paper. Note condition (c) in proposition 3 is included since by convention the Hausdorff distance between the empty set and any other set is $+\infty$. Finally note that proposition 3 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.

For an inference procedure for the parameters considered in this paper, note that each functional identified set can be seen as the projection of the identified set $\mathcal{Q}$ onto the real line. There are many existing procedures that could be used for uniformly valid confidence sets for the true joint distribution $Q \in \mathcal{Q}$ using the moment inequality characterization of $\mathcal{Q}$. Such procedures are typically based on inverting a hypothesis test of $H_0 : Q \in \mathcal{Q}$ versus an unrestricted alternative, where special care is taken when calculating the critical value to ensure uniform validity across a large class of data generating processes; see, for example, the approaches of Andrews and Guggenberger (2009) using subsampling; Andrews and Soares (2010) and Andrews and Shi (2014) using GMS critical values; and Beresteanu et al. (2011) in the context of models with convex moment predictions. In contrast, procedures that are uniformly valid for functionals of the identified set $\mathcal{Q}$ include Chernozhukov et al. (2015), Bugni et al. (2017), and Kaido et al. (2017a).[27]

For the environment in this paper, we find the approach of Kaido et al. (2017a) to be particularly promising, and we briefly describe their method here.[28] Assume for the purpose of inference that f is differentiable and take $\widetilde{Q} = Q + \lambda/\sqrt{n}$ to be a local expansion around Q for some $\lambda \in \mathbb{R}$. Note that — provided there are some conditions for the constraints imposed on $\mathcal{Q}^\dagger$, and after converting all moment equalities into two equivalent moment inequalities— the constraint set $\mathcal{Q}$ under Artstein's inequalities, or under the dual approach, can be written as a sequence of moment inequalities. Moreover, if the constraints on $\mathcal{Q}^\dagger$ are linear (as is the case with the MTR and MIV assumptions), then the constraint set $\mathcal{Q}$ can be characterized by a sequence of moment inequalities that are linear in the underlying distribution Q. We will see in a moment that this can yield significant computational advantages under the method of Kaido et al. (2017a).

To keep notation similar to that of Kaido et al. (2017a), we denote the sequence of population moment

[27] For a description of how to use the procedure in Chernozhukov et al. (2015) for a setting similar to the one in this paper, see the discussion in Torgovitsky (2016).

[28] While their method is mostly presented in terms of inference on a subvector of a partially identified parameter, they also mention briefly in the version Kaido et al. (2017a) that the procedure is valid for smooth functions of the partially identified parameter. Slightly more discussion of this extension to their paper is given in the conclusion of an earlier version of their paper, Kaido et al. (2016).

conditions defining $\mathcal{Q}$ as:

$$\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q)] \leq 0, \qquad h = 1, \ldots, H$$

In addition, define the estimates of the moments and their population standard deviations as:

$$\begin{aligned} \bar{m}_{n,h}(Q) &\equiv n^{-1} \sum_{i=1}^{n} m_h(W_i, Q), & h &= 1, \ldots, H \\ \hat{\sigma}_{n,j}(Q) &\equiv \left(n^{-1} \sum_{i=1}^{n} [m_h(W_i, Q) - \bar{m}_h(W_i, Q)]^2 \right)^{1/2}, & h &= 1, \ldots, H \end{aligned}$$

Here we have written the moments as unconditional moments, and thus excluded the dependence on an instrument. However, as discussed in an earlier version of their paper (Kaido et al. (2016)) the inference framework to be described is easily extended to the case with an instrument if the support of the instrument is finite.

The confidence set for a functional f of the joint distribution — denoted here as CS_n^f— is then defined in Kaido et al. (2017a) as:

$$CS_n^f = \left[\inf_{Q \in \mathcal{C}_n^f(\hat{c}_n)} f(Q), \sup_{Q \in \mathcal{C}_n^f(\hat{c}_n)} f(Q) \right]$$

where the constraint set $\mathcal{C}_n^f(\hat{c}_n)$ is defined as:

$$\mathcal{C}_n^f(\hat{c}_n) \equiv \left\{ Q : \quad n^{-1} m_h(W_i, Q)/\hat{\sigma}_{n,h}(Q) \leq \hat{c}_n(Q), \quad h = 1, \ldots, H \right\}$$

The intuition behind this construction is clear: the confidence set for the projection of the identified set under f is constructed through the minimization and maximization of f over a suitably relaxed version of the constraint set.[29] The degree of "relaxation" is governed by the parameter $\hat{c}_n(Q)$, which must be carefully calibrated in order to produce a confidence set with uniformly valid coverage over a large class of DGPs, while controlling for the level of projection conservatism. It should be no surprise that the computationally intense component of the procedure of Kaido et al. (2017a) is in the calibration of $\hat{c}_n(Q)$ over the suitably fine grid of the space of probability distributions Q. However, Kaido et al. (2017a) present an efficient algorithm to accomplish this step. Moreover, in the case when the moment inequalities are linear, we will see that are further computational gains.

To gain further insight into the procedure, note that by the definition of CS_n^f, the projection $f(Q)$ is covered when:

[29] Note that $\mathcal{C}_n^f(0) = \mathcal{Q}$.

$$\left\{\begin{array}{l} \inf\limits_{\widetilde{Q}} f(\widetilde{Q}) \\ s.t. \quad \widetilde{Q} \in \mathcal{Q}^{\dagger}, \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(\widetilde{Q})}{\hat{\sigma}_{n,h}(\widetilde{Q})} \leq \hat{c}_n(\widetilde{Q}) \quad \forall h \end{array}\right\} \leq f(Q) \leq \left\{\begin{array}{l} \sup\limits_{\widetilde{Q}} f(\widetilde{Q}) \\ s.t. \quad \widetilde{Q} \in \mathcal{Q}^{\dagger}, \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(\widetilde{Q})}{\hat{\sigma}_{n,h}(\widetilde{Q})} \leq \hat{c}_n(\widetilde{Q}) \quad \forall h \end{array}\right\} \tag{40}$$

$$\iff \left\{\begin{array}{l} \inf\limits_{\widetilde{Q}} f(Q+\lambda/\sqrt{n}) - f(Q) \\ s.t. \quad \widetilde{Q} \in \mathcal{Q}^{\dagger}, \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(\widetilde{Q})}{\hat{\sigma}_{n,h}(\widetilde{Q})} \leq \hat{c}_n(\widetilde{Q}) \quad \forall h \end{array}\right\} \leq 0 \leq \left\{\begin{array}{l} \sup\limits_{\widetilde{Q}} f(Q+\lambda/\sqrt{n}) - f(Q) \\ s.t. \quad \widetilde{Q} \in \mathcal{Q}^{\dagger}, \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(\widetilde{Q})}{\hat{\sigma}_{n,h}(\widetilde{Q})} \leq \hat{c}_n(\widetilde{Q}) \quad \forall h \end{array}\right\}$$

$$\iff \left\{\begin{array}{l} \inf\limits_{\lambda} \nabla_Q f(Q)\lambda \\ s.t. \quad \lambda \in \sqrt{n}(\mathcal{Q}^{\dagger} - Q), \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(Q+\lambda/\sqrt{n})}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})} \leq \hat{c}_n(Q+\lambda/\sqrt{n})) \quad \forall h \end{array}\right\}$$

$$\leq 0 \leq \left\{\begin{array}{l} \sup\limits_{\lambda} \nabla_Q f(Q)\lambda \\ s.t. \quad \lambda \in \sqrt{n}(\mathcal{Q}^{\dagger} - Q), \qquad \dfrac{\sqrt{n}\bar{m}_{n,h}(Q+\lambda/\sqrt{n})}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})} \leq \hat{c}_n(Q+\lambda/\sqrt{n}) \quad \forall h \end{array}\right\}$$

where the last step follows from taking a first-order Taylor-series expansion of f around $f(Q)$. Now note that by adding and subtracting $\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})]$ from the constraint in the constraint set we obtain:

$$\begin{aligned} \frac{\sqrt{n}\bar{m}_{n,h}(Q+\lambda/\sqrt{n})}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})} &= \sqrt{n}\frac{(\bar{m}_{n,h}(Q+\lambda/\sqrt{n}) - \mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})])}{\hat{\sigma}_{n,h}} + \sqrt{n}\frac{\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})]}{\hat{\sigma}_{n,h}} \\ &= \sqrt{n}\frac{(\bar{m}_{n,h}(Q+\lambda/\sqrt{n}) - \mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})])}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})} + \sqrt{n}\frac{\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})]}{\hat{\sigma}_{n,h}} \\ &= \frac{\sigma_{P,h}}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})}\left(\sqrt{n}\frac{(\bar{m}_{n,h}(Q+\lambda/\sqrt{n}) - \mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})])}{\sigma_{P,h}} + \sqrt{n}\frac{\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})]}{\sigma_{P,h}(Q+\lambda/\sqrt{n})}\right) \\ &= \frac{\sigma_{P,h}}{\hat{\sigma}_{n,h}(Q+\lambda/\sqrt{n})}\left(\mathbb{G}_{n,h}(Q+\lambda/\sqrt{n}) + \sqrt{n}\frac{\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})]}{\sigma_{P,h}(Q+\lambda/\sqrt{n})}\right) \end{aligned}$$

where $\mathbb{G}_{n,h}(\cdot)$ is a normalized, mean-zero empirical process. Now note that we can expand:

$$\mathbb{E}_{\mathbb{P}}[m_h(W_i, Q+\lambda/\sqrt{n})] = \mathbb{E}_{\mathbb{P}}[m_h(W_i, Q)] + \frac{1}{\sqrt{n}}\nabla_Q\mathbb{E}_{\mathbb{P}}[m_h(W_i, \bar{Q})] \cdot \lambda$$

where $\bar{Q} \in [\widetilde{Q}, \widetilde{Q}+\lambda/\sqrt{n}]$. Let $D_{\mathbb{P},h}(\bar{Q}) \equiv \nabla_Q\mathbb{E}_{\mathbb{P}}[m_h(W_i, \bar{Q})]$. Now note, in cases when $m_h(W_i, Q)$ is linear in Q, we will have that $D_{\mathbb{P},h}(\bar{Q}) = D_h$, which will simplify the calibration of $\hat{c}_n$.[30] Combining we have:

[30] The simplification follows from the fact that, with a fixed gradient, we do not need to compute the numerical gradient as we grid over the parameter space.

$$\frac{\sqrt{n}\bar{m}_{n,h}(Q+\lambda/\sqrt{n})}{\hat{\sigma}_{n,h}} = \frac{\sigma_{P,h}}{\hat{\sigma}_{n,h}}\left(\mathbb{G}_{n,h}(Q+\lambda/\sqrt{n}) + \frac{D_h \cdot \lambda}{\sigma_{P,h}} + \sqrt{n}\frac{\mathbb{E}_{\mathbb{P}}[m_h(W_i,Q)]}{\sigma_{P,h}}\right) \tag{41}$$

The final term in the brackets gives a measure of the slackness of the moment inequality constraints, which cannot be consistently estimated in a uniform sense. However, it can be conservatively approximated through the hard-threshold *generalized moment selection* (GMS) function proposed in Andrews and Soares (2010).

Based on the expansion of the constraints given in (41), and the discussion in Kaido et al. (2016), the critical value $\hat{c}_n(Q)$ will be chosen by

$$\hat{c}_n(Q) = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}^*(\Lambda_n^b(Q,\rho,c) \cap \{\nabla f(Q)\lambda = 0\} \neq \emptyset) \geq 1-\alpha\right\}$$

where $\mathbb{P}^*$ denotes the bootstrap distribution, and where

$$\Lambda_n^b(Q,\rho,c) = \left\{\lambda \in \rho B^d : \mathbb{G}_{n,h}^b(Q) + D_h\lambda + \varphi_h(\hat{\xi}_{n,h}(Q)) \leq c, \quad h = 1,2,\ldots,d_A\right\}$$

with

$$\mathbb{G}_{n,h}^b(Q) = \frac{1}{\sqrt{n}}\frac{\sum_{i=1}^n(m_{n,h}(W_i^b,Q) - \bar{m}_h(Q))}{\hat{\sigma}_{n,h}(Q)}$$

$$\hat{\xi}_{n,h}(Q) = \begin{cases} \kappa_n^{-1}\sqrt{n}\bar{m}_{n,h}(Q)/\hat{\sigma}_{n,h}(Q) & \text{if moment } h \text{ is not derived from a moment equality} \\ 0 & \text{if moment } h \text{ is derived from a moment equality} \end{cases}$$

$$\varphi_j(x) = \begin{cases} 0 & \text{if } x \geq -1 \\ -\infty & \text{if } x < -1 \end{cases}$$

and where $\rho > 0$ and κ_n are parameters chosen by the researcher and where

$$B^d = \{x \in \mathbb{R}^d : |x_h| \leq 1, \forall h\}$$

is the unit box in $\mathbb{R}^d$.[31] See the discussion in Kaido et al. (2017a) for how to choose the tuning parameters.

When the dimension of Q is large, it can be computationally intensive to grid over the space $\mathcal{Q}^\dagger$. To

[31] I.e. the critical level $\hat{c}_n(Q)$ is then the smallest value of c that makes the bootstrap probability of the event

$$\min_{\lambda \in \Lambda_n^b(Q,\rho,c)} \nabla_Q f(Q)\lambda \leq 0 \leq \max_{\lambda \in \Lambda_n^b(Q,\rho,c)} \nabla_Q f(Q)\lambda$$

at least $1-\alpha$.

deal with the computational burden of the approach, Kaido et al. (2017a) propose using a response surface method to efficiently optimize over the parameter space. This involves approximating the functions $f(Q)$ and $\hat{c}_n^A(Q)$ using a flexible auxiliary model.[32] Following the notation in Kaido et al. (2017a), let $\hat{c}_n^A(Q)$ be the flexible auxiliary model estimated on a grid with L points that grows linearly with the dimension of the parameter space. Using the auxiliary model we can easily compute $\nabla_\theta \hat{c}_n^A(Q)$ which may help increase the speed of computation. After estimating $\hat{c}_n^A(Q)$ using suitably many grid points, we can then run the following program:

$$\max / \min_{Q \in \mathcal{Q}^\dagger} f(Q) \qquad s.t. \qquad \sqrt{n}\bar{m}_{n,h}(Q)/\hat{\sigma}_{n,h}(Q) \leq \hat{c}_n^A(Q) \tag{42}$$

After optimal values are found, we can draw additional points from a subset of the parameter space around the points that obtain the optimum and add them to the existing points, reconstruct $\hat{c}_n^A(Q)$ and repeat the steps above. The confidence interval is then constructed when the program converges to a stable maximum and minimum value in (42).

We leave it up to the researcher to verify the assumptions of Kaido et al. (2017a) in their particular application. However, under their assumptions, we have by (slight modification of) theorem 4.1 in Kaido et al. (2017a) that:

$$\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathscr{P}} \inf_{Q\in\mathcal{Q}^\dagger} \mathbb{P}(f(Q) \in CS_n^f) \geq 1-\alpha$$

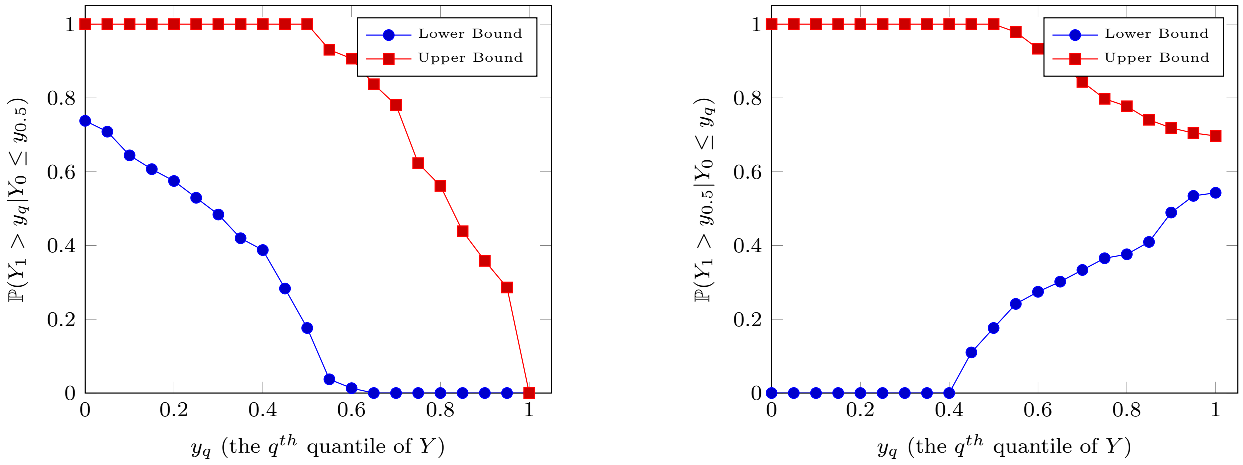
which establishes the the confidence set CS_n^f is uniformly asymptotically valid over the class of DGPs $\mathscr{P}$ defined by the assumptions of Kaido et al. (2017a).

Software for this approach is now free available online from `https://molinari.economics.cornell.edu/programs.html`, with a corresponding user guide provided in Kaido et al. (2017b).

[32] In Kaido et al. (2017a) this is done using Gaussian Process regression.

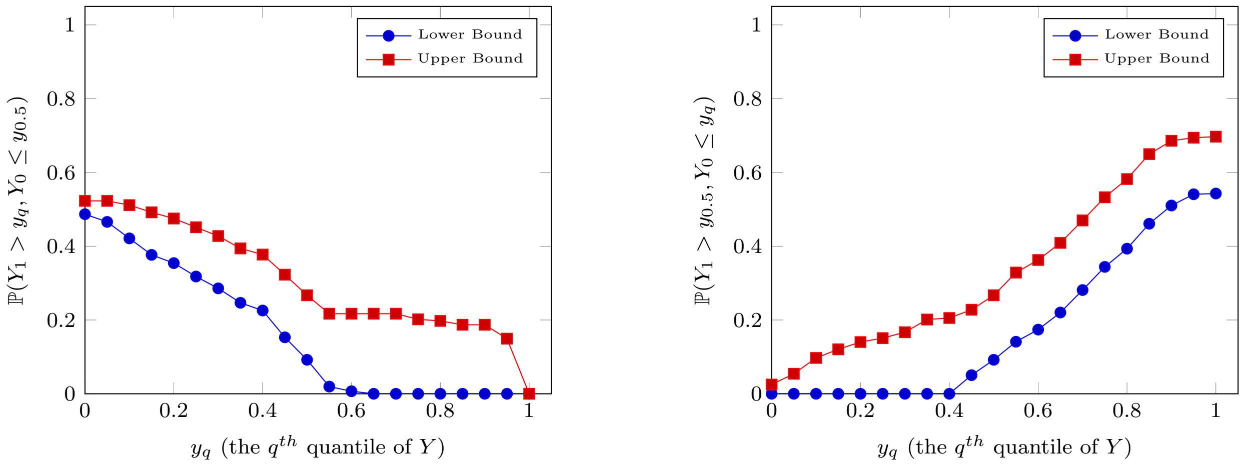
Appendix E Application Robustness Exercise

Figure 3 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 4 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.



(a) *(b)*

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



(a) *(b)*

Figure 4: Bounds on the joint probability (Grade 3, Bins=35, MTR assumption $\mathbb{P}(Y_1 > Y_0) \geq 0.95$).

Table 5 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 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 will always deliver sharp bounds).

Table 5: 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	0.07	0.50	0.02	0.48
	Bins=30	0.08	0.53	0.04	0.52
	Bins=35	0.09	0.53	0.04	0.51
$\mathbb{E}[Y_1 - Y_0]$:	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
$\mathbb{P}(Y_1 > Median(Y) \mid Y_0 \leq Median(Y))$:†	Bins=25	0.14	1.00	0.05	1.00
	Bins=30	0.15	1.00	0.07	0.99
	Bins=35	0.16	0.97	0.08	0.95
$\mathbb{P}(Y_0 \leq Median(Y))$:†	Bins=25	0.58	0.58	0.59	0.59
	Bins=30	0.55	0.55	0.52	0.52
	Bins=35	0.53	0.53	0.51	0.51
$\mathbb{P}(Y_1 > Median(Y))$:†	Bins=25	0.36	0.62	0.32	0.59
	Bins=30	0.39	0.66	0.39	0.65
	Bins=35	0.42	0.69	0.40	0.67
$Corr(Y_0, Y_1)$:	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 $\mathbb{P}(Y_1 > Y_0) \geq 0.95$.

Appendix F Proofs

Proof of Theorem 1. If $\mathcal{U}$ is finite, then so is $G^{-1}(Y,D)$ since G^{-1} maps within $\mathcal{U}$. Thus, $\{(Y,D) : G^{-1}(Y,D) \cap A \neq \emptyset\} \in 2^{\mathcal{W}}$ 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, say with dimension d_u, we have that $\mathcal{Q} \subset \mathbb{R}^{d_u}$ is compact with respect to the usual topology with the euclidean norm. Finally, the image of a continuous functional over a non-empty compact and convex set $\mathcal{Q} \subset \mathbb{R}^{d_u}$ is a nonempty interval with the end points defined as in (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 will 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 1 in definition 7 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_2, and shares the last element with vector u_1. 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:

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

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 $u \in \mathcal{U}$ (this can be proven by inducting on the dimension of $\mathcal{Y}$). 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 7). 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 an 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 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 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 will 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 $\mathcal{G}_k = (\mathcal{Y}_k, \mathcal{U}, G)$ and where

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

By construction we have $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for all $i \neq j$, $G^{-1}(y) \cap G^{-1}(y') = \emptyset$ and for any pair $(y,y') \in \mathcal{Y}_k$. 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 $G^{-1}(\mathcal{Y}_k) = \mathcal{U}$. Finally consider the pair (v,v') with $v \in \mathcal{Y}_i$, $v' \in \mathcal{Y}_j$, $i \neq j$. 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 the property of collections $\mathcal{Y}_k$ proved in part (a), (i) is implied if B is connected. In addition, note that $G^{-1}(\mathcal{Y}_k) = \mathcal{U}$ 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 connected. Next, suppose by way of contradiction that there exists a collection of singletons $B = \{y_1, \ldots, y_r\} \subset \mathcal{W}$ satisfying the above conditions, 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, but since B does not contain v this is only possible if $\mathcal{Y}_k \subseteq B$ 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

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

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

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

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

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 $B \subseteq \mathcal{Y}_2$ 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 $B \nsubseteq \mathcal{Y}_k$ for any k and $\mathcal{Y}_k \nsubseteq B$ 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 $\mathcal{Y}_2, \ldots, \mathcal{Y}_k$ regardless of the magnitude of r, and $\mathcal{Y}_k \nsubseteq B$ 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 Proposition 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 7, 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 $\mathcal{Y}_k$, 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, then we must 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 Proposition 2. Suppose first that Q is known up to a finite-dimensional parameter $\theta \in \Theta$ where Θ is compact. Denote the dependence of Q on θ as $Q(\cdot|\theta) = Q_\theta(\cdot)$. Identical to Galichon and Henry (2011), define the identified set for θ as:

$$\Theta_I = \left\{ \theta \in \Theta : Q_\theta(A) \leq P(G^{-1}(Y,D) \cap A \neq \emptyset) \qquad \forall A \in 2^{\mathcal{U}} \right\}$$

I.e. the identified set is the collection of θ's such that the distribution $Q(\cdot|\theta)$ satisfies Artstein's inequalities. By theorem 1 and theorem 3 in Galichon and Henry (2011) we know that for some $\widetilde{\theta} \in \Theta$, we have that $\widetilde{\theta} \in \Theta_I$ if and only if there exists a joint distribution $\pi \in \mathcal{M}_G(P, Q_\theta)$.

Since $\mathcal{U} = \mathcal{Y}^K$ is assumed to be finite, take $\Theta = \mathcal{Q}^\dagger$ and $\theta = Q$ (since any $Q \in \mathcal{Q}^\dagger$ is a finite dimensional vector). Then by theorem 1 and 3 in Galichon and Henry (2011) we have that

$$\left\{ \theta \in \Theta : Q_\theta(A) \leq P(G^{-1}(Y,D) \cap A \neq \emptyset) \quad \forall A \in 2^{\mathcal{U}} \right\} = \{ \theta \in \Theta : \exists \pi \in \mathcal{M}_G(P, Q_\theta) \}$$

$$\iff \left\{ Q \in \mathcal{Q}^\dagger : Q(A) \leq P(G^{-1}(Y,D) \cap A \neq \emptyset) \quad \forall A \in 2^{\mathcal{U}} \right\} = \left\{ Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}_G(P, Q) \right\}$$

Thus we have that for each fixed $Z = z$:

$$\begin{aligned}\mathcal{Q}_z &= \left\{Q \in \mathcal{Q}^\dagger : Q(A) \leq P(G^{-1}(Y,D) \cap A \neq \emptyset | Z = z) \quad \forall A \in 2^{\mathcal{U}}\right\} \\ &= \left\{Q \in \mathcal{Q}^\dagger : \exists \pi \in \mathcal{M}_{G|z}(P_z, Q)\right\} \\ &= \mathcal{Q}_z^*\end{aligned}$$

■

Proof of Proposition 3. First we will show that $\hat{\Theta}^f$ is indeed a random set, which requires showing that it is an Effros-measurable map. This is needed to show that convergence in the Hausdorff metric is well-defined (see Kaido and White (2014)). Consider our probability space $(\Omega, \mathbb{P}, \mathcal{F})$. Let $\mathcal{F}(\mathcal{A})$ represent the closed subsets of a compact subset $\mathcal{A} \subset \mathbb{R}^d$. Note a map $\mathbf{X} : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F}(\mathcal{A}))$ is Effros-measurable with respect to $\mathcal{F}$ if:

$$\mathbf{X}^{-1}(F) = \{\omega : g(\omega) \cap F\} \in \mathcal{F} \qquad \forall F \in \mathcal{F}(\mathcal{A})$$

Now note that $\hat{\Theta}^f = f(\hat{\mathcal{Q}}_n)$, so that if we can show $\hat{\mathcal{Q}}_n$ is Effros-measurable then we are done. Indeed, by continuity of f, for any closed set B we have $A \equiv f^{-1}(B)$ is also closed so that

$$[\hat{\Theta}^f]^{-1}(B) = \hat{\mathcal{Q}}_n^{-1}(f^{-1}(B)) = \hat{\mathcal{Q}}_n^{-1}(A) = \{\omega : \hat{\mathcal{Q}}_n(\omega) \cap A\} \in \mathcal{F}$$

where the last line follows if we can show Effros-measurability of $\hat{\mathcal{Q}}_n = \hat{\mathcal{Q}}_n(\omega, \hat{P}_n)$. Let Δ_Q and Δ_P denote the probability simplexes where $\hat{\mathcal{Q}}_n$ and $\hat{P}_n$ take their respective values; i.e. $\hat{\mathcal{Q}}_n : \Delta_P \rightrightarrows \Delta_Q$. Clearly Δ_Q and Δ_P are both compact. In addition, later in the proof we will show $\hat{\mathcal{Q}}_n(\omega, \hat{P}_n)$ can be reformulated as the level set of a random criterion function that is a continuous function on Δ_P. Note that this environment is the same as that considered in Kaido and White (2014) in the special case when the first-stage parameter $(\hat{P}_n)$ is point-identified, and so Effros-measurability of $\hat{Q}_n$ will follow from their theorem 4.1. This confirms that $\hat{\Theta}^f$ is Effros-measurable (i.e. a random closed set) and that convergence in the Hausdorff metric is well-defined.

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 $\hat{f}_n^u \xrightarrow{p} f^u$. Since $\mathcal{U}$ is finite and $\mathcal{Q}^\dagger$ is closed and compact, $\mathcal{Q}$ is convex and compact with respect to the euclidean norm. Since f is a continuous functional, and since the inf and sup of a continuous functional over a compact set is a continuous functional, the result follows from the continuous mapping theorem if we can show that $d_H(\hat{\mathcal{Q}}_n, \mathcal{Q}) \xrightarrow{p} 0$. This can be proven using the consistency result of Shi and Shum (2015).

Note that there is a $(d_A + d_{\mathcal{U}}) \times 1$ vector of moment conditions implied by Artstein's inequalities and the non-negativity constraints on Q, where d_A denotes the number of inequalities defined by Artstein's theorem. We can convert these inequalities to equalities by introducing a $(d_A + d_{\mathcal{U}}) \times 1$ slackness parameter $\lambda \geq 0$. Let $\theta = (Q, \lambda)$ denote the $(d_\theta \times 1)$ parameter vector of interest, where $d_\theta = d_A + d_{\mathcal{U}}$, and let $g(\theta, P)$ be the $d_\theta \times 1$ 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 $g_k(\theta, W_i)$, and the associated slackness parameter λ_k, corresponding to the singleton $u_k \in U$, solve for λ_k using the constraint:

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

$$\implies \lambda_k = 1 - \sum_{j \in I(\mathcal{U})} Q(U = u_j) - \sum_{j \in I(\mathcal{U}), j \neq k} \lambda_j \tag{43}$$

and then add the non-negativity constraint on (43) (where $I(\mathcal{U})$ is an index set for elements in $\mathcal{U}$). Thus, there will be $(d_\theta - 1) \times 1$ equality constraints $g(\theta, P)$, and $d_\theta \times 1$ inequality constraints given by

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

Importantly, the inequality constraints do not depend on the first-stage parameter P. Now define $\Theta_0 = \{\theta \in \Theta : g(\theta, P) = 0, h(\theta) \geq 0\}$. Consider the criterion function:

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

Then we have:

$$\Theta_0 = \arg\min_{\theta \in \Theta} T(\theta, P) \quad s.t. \quad h(\theta) \succcurlyeq \mathbf{0}$$

The sample analog of the above is:

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

Thus to show $d_H(\hat{\mathcal{Q}}_n, \mathcal{Q}) \xrightarrow{p} 0$ it suffices to show $d_H(\hat{\Theta}_n, \Theta_0) \xrightarrow{p} 0$. To do this, it suffices to verify the conditions of theorem 2.1 in Shi and Shum (2015):

1. By Glivenko-Cantelli we know that $\hat{P}_{n,z} \to P_z$ uniformly in probability for all z.
2. The space of observable distributions $\mathcal{P}$ is compact. Θ is also compact (since it is without loss of generality that we restrict $\lambda \in [0, 1]$; the moment conditions will ensure this is true).

3. $g(\cdot, P)$ is trivially continuously differentiable on Θ for all $P_z \in \mathcal{P}$ and all z, and $h(\cdot)$ is trivially continuous on Θ (both are linear functions of Θ).

4. Θ_0 is defined completely by linear equality and inequality constraints, it is closed and convex, which implies $cl(int(\Theta_0)) = \Theta_0$. In addition we have have $(d_\theta - 1) \times 1$ equality constraints, d_θ unknown parameters, and the Jacobian $\partial g(\theta, P)/\partial \theta'$ has full row rank.

Consistency of $\hat{\Theta}_n$ for Θ_0 in the Hausdorff metric then follows from theorem 2.1 in Shi and Shum (2015).

■