Identification and Inference for Functionals of Partially Identified Parameters

by

Thomas M. Russell

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Economics
University of Toronto

Abstract

Identification and Inference for Functionals of Partially Identified Parameters

Thomas M. Russell
Doctor of Philosophy
Graduate Department of Economics
University of Toronto
2020

This thesis presents three essays related to identification and inference for functionals of partially identified parameters. This work is motivated by the fact that often the final object of interest in an empirical setting is not the whole vector of structural parameters, but is instead a single element of the parameter vector or, more generally, a functional of the parameter vector. A common theme throughout is that bounds on functionals of partially identified parameters can often be formulated as the solution to two stochastic optimization problems, one for the upper bound and one for the lower bound. In the first essay, we introduce an identification and estimation result for bounding functionals of the joint distribution of potential outcomes from the literature on treatment effects, and present an application to the evaluation of class sizes on test scores. The second essay considers the problem of inference for functionals of partially identified parameters defined in terms of stochastic linear programs. Under some regularity conditions, it is shown that a naive bootstrap procedure can be used to construct uniformly valid confidence sets for the true value of the partially identified functional of interest. In the final essay, the problem of counterfactuals and policy choice is considered. It is shown that a specific class of counterfactual objects, called policy transforms, can be bounded as the solution to two optimization problems without the need to impose parametric distributional assumptions on the latent variables in the model. The notion of learnability of optimal policies is defined, and sufficient conditions are provided for a class of policies to be learnable. Finally, finite sample theoretical guarantees for certain policy rules are derived. All chapters in this thesis are self-contained.

To my wife, Alexis.

Acknowledgements

Chapter 1: As per the copyright agreement with the American Statistical Association (ASA), the author acknowledges the original publication of this chapter in the Journal of Business and Economic Statistics in 2019.[1] I am grateful to the editor, an associate editor, and to three anonymous referees for excellent feedback. I am indebted to Ismael Mourifié for countless hours of discussion that have helped to improve this chapter. I am also especially grateful to Victor Aguirregabiria, JoonHwan Cho, Christian Gourieroux, Jiaying Gu, and Yuanyuan Wan who have all provided valuable feedback, and to Marc Henry for hosting me at the Penn State Department of Economics. The chapter benefited from conversations with Stephane Bonhomme, Sung Jae Jun, Desiré Kedagni, Toru Kitagawa, Joris Pinske, and Adam Rosen. I am also grateful to the 2018 International Association of Applied Econometrics (IAAE) conference organizers for the IAAE travel grant that made it possible to present the results from this chapter.

Chapter 2: This chapter was written jointly with JoonHwan Cho, a PhD candidate at the University of Toronto. A previous version of this chapter was circulated as a paper under the title "Inference on Functionals of Set-Identified Parameters Defined by Convex Moments." We are grateful to Victor Aguirregabiria, Bulat Gafarov, Christian Gourieroux, Jiaying Gu, Ismael Mourifie, Jeffrey Negrea, Brennan Thompson, Stanislav Volgushev and Yuanyuan Wan for helpful comments and discussion. We are also grateful to participants at the 7th Annual Doctoral Workshop in Applied Econometrics at the University of Toronto, as well as participants at the 2019 North America Summer Meeting of the Econometric Society at the University of Washington.

Chapter 3: I thank Jiaying Gu, Ismael Mourifie, Eduardo Souza-Rodrigues, Stanislav Volgushev and Yuanyuan Wan for feedback and encouragement, and I am especially grateful to JoonHwan Cho for many hours of discussion that helped to improve this paper.

Finally, I am grateful to Adam Rosen for taking the time to serve as an external member of my Thesis committee, and for his helpful sugggestions and encouraging remarks.

All the research in this thesis was supported by the Social Sciences and Humanities Research Council of Canada.

[1] https://doi.org/10.1080/07350015.2019.1684300

Contents

Chapter 1

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

This chapter proposes an identification and estimation method that allows researchers to bound continuous functionals of the joint distribution of potential outcomes from the literature on treatment effects. The focus is on a model where no restrictions are imposed on treatment selection. The method can sharply bound interesting parameters when analytical bounds are 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.

1.1 Introduction

This chapter 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 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 chapter 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. 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.

The method accomplishes 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. 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 be solved as a minimization (for the lower bound) and maximization (for the upper bound) problem subject to a carefully selected set of constraints. Despite the fact that the bounds are not represented analytically, our proposed procedure ensures that they are still sharp.[1] To bound a continuous functional of the joint distribution using an optimization approach 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 (2017a), among others, and is also used in the main identification results for this chapter. The second characterization re-frames the bounding problem as an optimal transport problem, and has been considered in Galichon and Henry (2011) and Lafférs (2013b, 2015). The two characterizations are compared based on their computational tractability, and the conditions under which one approach dominates the other are discussed. A result is then presented that shows a researcher can obtain a consistent estimate of the identified set using either approach.

Finally, we apply the theoretical results to data from the Tennessee STAR experiment considered in Krueger (1999) and Krueger and Whitmore (2001). As described in Krueger (1999), the Tennessee STAR experiment saw students randomized into small and large classrooms with the goal of evaluating the impact of class size on test scores. However, the experiment was affected by imperfect compliance with treatment assignment. Using our approach, we find that bounds on the average treatment effect 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 modelling assumptions.

This chapter extends work by Galichon and Henry (2006, 2009, 2011), Beresteanu and Molinari (2008), Beresteanu et al. (2011, 2012), and Chesher and Rosen (2017a). 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 chapter also serves as a framework to facilitate the introduction of additional model assumptions and structure.[2]

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 pre-

[1] Bounds on a parameter are called *sharp* if they are the smallest bounds consistent with the observed data and the researcher's model assumptions.

[2] The approach is also in the spirit of Ginther (2000), who shows estimates of returns to schooling are sensitive 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.

sented in Chiburis (2010), Lafférs (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). Finally, this chapter is also related to the papers of Fan et al. (2017) and Firpo and Ridder (2019). However, unlike Fan et al. (2017) and Firpo and Ridder (2019) our method does not require that the marginal distributions of potential outcomes be identified. Furthermore, Firpo and Ridder (2019) focus specifically on bounds on functionals of the distribution of treatment effects (i.e. the difference between potential outcomes), and not bounds on functionals of the joint distribution of potential outcomes, which are considered here.

1.2 Identification

1.2.1 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 loss of generality we take $\mathcal{D} = \{0, 1, \ldots, K-1\}$. When possible, we use the notation $W := (Y, D)$ and $\mathcal{W} := \mathcal{Y} \times \mathcal{D}$. With this defined, we assume throughout that $\mathcal{W}$ is a finite subset of a 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 $U := (Y_0, Y_1, \ldots, Y_{K-1}) \in \mathcal{U}$, 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.[3] Finally, we denote the $\sigma-$algebra on U as $2^{\mathcal{U}}$ and we refer to the vector of random variables U as *potential outcomes*.

All random variables in this chapter are assumed to be defined on a probability space $(\Omega, \mathfrak{A}, P)$. Let P_W denote the distribution induced on $\mathcal{W}$ by (Y, D), and let P_U denote the distribution induced by U on $\mathcal{U}$. In particular:

$$P_W(A) = P(\omega : W(\omega) \in A), \qquad A \in 2^{\mathcal{W}},$$
$$P_U(B) = P(\omega : U(\omega) \in B), \qquad B \in 2^{\mathcal{U}}.$$

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

Definition 1.2.1 (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 $K \geq 2$.

Importantly, note that the mechanism that determines the selection variable D has not been specified; we return to this point in the next subsection.

The objective of this chapter is to recover the distribution P_U, and functionals thereof, using the observed distribution of (Y, D). In this section we assume that the researcher has knowledge of the distribution P_W. 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.

[3] 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.

Given 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. For now we assume the instrument Z is a random variable that affects the treatment choice D but is 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 let $P_{W|Z}$ denote the conditional distribution induced on $\mathcal{W}$ by (Y, D) given a generic value of $Z = z$. In particular:

$$P_{W|Z}(A) = P(\omega : W(\omega) \in A | Z = z), \qquad A \in 2^{\mathcal{W}}.$$

In this chapter we leave the dependence between Z and D completely unrestricted.

With the preliminaries in place, we provide some discussion of functionals f of the joint distribution P_U. Heckman et al. (1997) argue that there are many interesting parameters in the program evaluation literature that require knowledge of the joint distribution. One trivial example is to take f to represent the average treatment effect between treatments $D = d$ and $D = d'$ for $d, d' \in \mathcal{D}$:

$$f(P_U) = \int_{\mathcal{U}} (Y_{d'} - Y_d) \, dP_U.$$

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

(i) *The correlation between potential outcomes:*

$$f(P_U) = \frac{\int_{\mathcal{U}} (Y_d - \mathbb{E}(Y_d)) (Y_{d'} - \mathbb{E}(Y_{d'})) \, dP_U}{\left(\int_{\mathcal{U}} (Y_d - \mathbb{E}(Y_d))^2 \, dP_U \right)^{1/2} \left(\int_{\mathcal{U}} (Y_{d'} - \mathbb{E}(Y_{d'}))^2 \, dP_U \right)^{1/2}}.$$

This parameter provides a simple measure of the dependence of outcomes across the states $D = d$ and $D = d'$. 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.[4] 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 of Y_d and $Y_{d'}$, 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:*

$$f(P_U) = \int_{\mathcal{U}} \mathbb{1}\{Y_{d'} > Y_d\} \, dP_U.$$

This parameter provides a measure of the proportion of individuals who benefit from treatment d' versus treatment d. 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).

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

(iii) *Distributional Mobility:*

$$f(P_U) = P_U(Y_{d'} \in A | Y_d \in B) = \frac{\int \mathbb{1}\{Y_{d'} \in A, Y_d \in B\}\, dP_U}{\int \mathbb{1}\{Y_d \in B\}\, dP_U}.$$

This parameter measures the probability that treatment helps an individual obtain an outcome $Y_{d'} \in A$, given his/her outcome under treatment d is fixed at $Y_d \in B$. 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:*

$$f(P_U) = \int_{\mathcal{U}} \left((Y_{d'} - Y_d) - \mathbb{E}(Y_{d'} - Y_d)\right)^2 \, dP_U.$$

The variance of treatment effects can provide a measure of the heterogeneity of treatment effects. If the potential outcomes $Y_{d'}$ and Y_d are dependent, then this parameter requires knowledge of the joint distribution P_U.

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 P_U that are consistent with the observed distribution P_W and the researcher's assumptions. Denote this set as $\mathcal{P}_U$. Under the assumption that $\mathcal{U}$ is finite, this set is convex and compact with respect to the euclidean norm. Next, we bound any continuous function $f : \mathcal{P}_U \to \mathbb{R}$ by noting that the image of a continuous function over a compact set is an interval $[f^\ell, f^u]$ where:

$$f^u = \sup_{P_U \in \mathcal{P}_U} f(P_U), \qquad f^\ell = \inf_{P_U \in \mathcal{P}_U} f(P_U). \tag{1.1}$$

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

1.2.2 Identification Without an Instrument

The formal identification argument applies results from random set theory.[5] This section reviews some of the concepts provided in Beresteanu et al. (2012) on the application of random set theory to the POM, and then provides a result that allows researchers to easily bound functionals of the joint distribution.

Let $\mathcal{P}_U^\dagger$ be the collection of all admissible distributions, and let $\boldsymbol{G}$ denote a model correspondence $\boldsymbol{G} : \mathcal{U} \to \mathcal{Y}$ mapping unobservables to observables. The set of admissible distributions $\mathcal{P}_U^\dagger$ represents the set of all distributions that satisfy the researcher's *a priori* restrictions on the distribution of U.[6] Recall that $\mathcal{P}_U$ represents the collection of distributions that satisfy the researcher's restrictions *and* that are consistent with the observed distribution P_W, so we have $\mathcal{P}_U \subseteq \mathcal{P}_U^\dagger$. In the absence of any additional restrictions, we

[5] 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 (2017a), among others.

[6] For example, we may take $\mathcal{P}_U^\dagger$ 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 1.C. Alternatively, in the presence of an instrument, we might consider the independence, mean independence, and quantile independence conditions discussed in Chesher and Rosen (2017a).

can simply take $\mathcal{P}_U^\dagger$ to be the $(|\mathcal{U}|-1)$-dimensional probability simplex. The model correspondence for the POM is given by:

$$\boldsymbol{G}(y_0, y_1, \ldots, y_{K-1}) = \left\{ (y,d) \in \mathcal{W} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d=k\} \right\}, \tag{1.2}$$

for $(y_0, y_1, \ldots, y_{K-1}) \in \mathcal{U}$. Following the definition provided, for example, in Tamer (2003), Lewbel (2007), Chesher and Rosen (2012) and Chesher and Rosen (2017a), a model is *incomplete* if, given values of the latent and exogenous variables, it is not possible to uniquely determine the values of the outcome variables. Applying this definition, the POM described so far is *incomplete*, namely since the mechanism generating the (possibly endogenous) variable D has not been specified. In other words, given values of the latent potential outcomes $(Y_0, \ldots, Y_{K-1})$, it is not possible to uniquely determine the values of the outcome variables (Y, D) since the mechanism generating D is left completely unrestricted. The result of model incompleteness is that the mapping $\boldsymbol{G} : \mathcal{U} \to \mathcal{W}$ is given by a correspondence rather than a function.[7]

There are two obvious reasons why a researcher might refrain from specifying the selection equation, and thus opt to work with the incomplete POM. First, any unobserved heterogeneity affecting choices and/or outcomes is left completely unrestricted.[8] 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 chapter under minor modifications.[9]

Taking the incomplete nature of this model as given, we can equivalently focus on learning the distribution P_U through the reverse correspondence:

$$\boldsymbol{G}^{-1}(y,d) = \left\{ (y_0, y_1, \ldots, y_{K-1}) \in \mathcal{U} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d=k\} \right\}. \tag{1.3}$$

Upon closer inspection, note the correspondence $\boldsymbol{G}^{-1}(y,d)$ will map (y,d) to the all vectors $(y_0, y_1, \ldots, y_{K-1})$ with $y_d = y$ and with all other elements taking arbitrary values in $\mathcal{Y}$. Following the discussion in Beresteanu et al. (2012), $\boldsymbol{G}^{-1} : \mathcal{W} \to \mathcal{U}$ is a *random closed set* (see Appendix 1.A 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 there exists a random variable $\widetilde{U}^*$ and a random vector (Y^*, D^*) such that $\widetilde{U}^* \sim \widetilde{U}$, $(Y^*, D^*) \sim (Y, D)$, and $\widetilde{U}^* \in \boldsymbol{G}^{-1}(Y^*, D^*)$ a.s. In this case, $\widetilde{U}^*$ is called a *selection* from the random closed set $\boldsymbol{G}^{-1}(Y^*, D^*)$. 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_W given the correspondence in (1.2). In such models, the observed distribution characterizes the random set $\boldsymbol{G}^{-1}(Y,D)$ through the *generalized likelihood*:

$$T(A) := P_W(\boldsymbol{G}^{-1}(Y,D) \cap A \neq \emptyset), \tag{1.4}$$

defined for every $A \in 2^{\mathcal{U}}$; see the discussion in Galichon and Henry (2011). The functional T is sometimes called the *capacity functional* of the *random set* $\boldsymbol{G}^{-1}(Y,D)$ (see Appendix 1.A for a formal definition). Note that, given $\boldsymbol{G} : \mathcal{U} \to \mathcal{W}$ is also a random set, we could have also defined a capacity functional for the random

[7] For an especially clear discussion of the distinction between complete and incomplete models, see Chesher and Rosen (2012).

[8] 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.

[9] 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.

set $\boldsymbol{G}(U)$ as:

$$T_{\mathcal{W}}(A) := P_U(\boldsymbol{G}(U) \cap A \neq \emptyset), \tag{1.5}$$

for every $A \in 2^{\mathcal{W}}$. We say that the capacity functional given in (1.4) has been defined on the *observables*, whereas the capacity functional given in (1.5) 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 (2017a) 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 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 $P_W = (p_{00}, p_{10}, p_{01}, p_{11})$ *and* $P_U = (q_{00}, q_{10}, q_{01}, q_{11})$ *where*

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

The reverse correspondence $\boldsymbol{G}^{-1} : \mathcal{Y} \to \mathcal{U}$ *for this model is defined by:*

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

The capacity functional associated with this reverse correspondence is given by (1.4)*, taking* A *to be any subset of* $\mathcal{U}$. *For example, if* $A = \{(0,0),(0,1)\}$*, then the capacity functional is given as:*

$$T(A) = T(\{(0,0),(0,1)\}) = P_W(\boldsymbol{G}(Y,D)^{-1} \cap \{(0,0),(0,1)\} \neq \emptyset) = p_{00} + p_{01} + p_{11}. \tag{1.6}$$

Given the capacity functional of the random set $\boldsymbol{G}^{-1}(Y,D)$, it is possible to characterize $\mathcal{P}_U$, the set of all distributions $P_U \in \mathcal{P}_U^{\dagger}$ that are consistent with the observed distribution P_W. As discussed in Beresteanu et al. (2012), a result from random set theory called *Artstein's Theorem* (Artstein (1983))—stated formally in Appendix 1.A—provides us with the necessary and sufficient conditions for the existence of a random variable $\widetilde{U}$ with distribution $P_U \in \mathcal{P}_U^{\dagger}$ that can rationalize the observed distribution P_W through the model correspondence $\boldsymbol{G}$. The necessary and sufficient conditions provided by Artstein's Theorem are captured by a set of inequality constraints on the distribution P_U:

$$P_U(A) \leq T(A), \qquad \forall A \in 2^{\mathcal{U}}.$$

This implies that we can write our collection $\mathcal{P}_U$—the collection of $P_U \in \mathcal{P}_U^{\dagger}$ that are consistent with the observed distribution P_W—as:

$$\mathcal{P}_U = \{P_U \in \mathcal{P}_U^{\dagger} : P_U(A) \leq T(A) \text{ for all } A \in 2^{\mathcal{U}}\}. \tag{1.7}$$

Since $\mathcal{U}$ is a finite set, to verify that a given distribution P_U can rationalize the observed distribution

P_W requires the researcher to check if a finite number of linear inequality constraints are satisfied. If the constraints defining $\mathcal{P}_U^\dagger$ are also linear, then $\mathcal{P}_U$ is a polyhedron contained in the $(|\mathcal{U}|-1)$-dimensional probability simplex. See Figure 1.1 for a visual representation of Artstein's inequalities in the setting of Example 1.

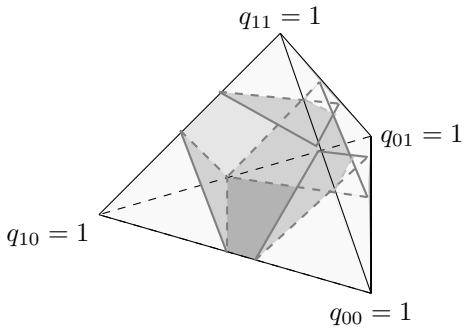


Figure 1.1: A representation of the constraints imposed by Artstein's Theorem on the probability simplex in the case when $Y, D \in \{0,1\}$. Here, the triangular pyramid represents the probability 3-simplex, and the shaded region within shows the distributions satisfying all of Artstein's inequalities.

Example 1 (Cont'd)**.** *Consider again the case where $Y \in \{0,1\}$ and $D \in \{0,1\}$. By Artstein's Theorem, the joint distribution P_U must respect the inequality constraint $P_U(A) \leq T(A)$ for every subset A of $\mathcal{U}$. For example, taking $A = \{(0,0),(0,1)\}$, Artstein's Theorem implies that P_U must satisfy:*

$$P_U(A) \leq T(A), \qquad \iff \qquad q_{00} + q_{01} \leq p_{00} + p_{01} + p_{11}. \tag{1.8}$$

Note that there are $2^{|\mathcal{U}|} = 16$ subsets of $\mathcal{U}$, and thus Artstein's Theorem implies that there are 16 such linear inequality constraints that must be satisfied by P_U for it to be consistent with the observed distribution P_W.

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

Theorem 1.2.1. *Let $\mathcal{P}_U^\dagger$ be a convex set of admissible distributions. If $\mathcal{P}_U$ is non-empty, then for every continuous functional $f : \mathcal{P}_U \to \mathbb{R}$, the identified set for f is a non-empty interval $[f^\ell, f^u]$ where:*

$$f^u = \sup_{P_U \in \mathcal{P}_U} f(P_U), \qquad f^\ell = \inf_{P_U \in \mathcal{P}_U} f(P_U). \tag{1.9}$$

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

Although the Theorem is stated for the case when $\boldsymbol{G}$ is a correspondence—so as to accommodate the POM—it applies equally to the case when $\boldsymbol{G}$ 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 (1.4) on the observables, the Theorem could have been written in an analogous manner using the capacity functional (1.5) defined on the unobservables.

Note that Theorem 1.2.1 is of interest from a practical point of view since—although it is an identification result—it suggests a straightforward method of estimation. Indeed, since the restrictions that define $\mathcal{P}_U$ 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. 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.[10]

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.[11] In contrast, Theorem 1.2.1 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 $\mathcal{P}_U^\dagger$, the results also extend easily to accommodate additional modelling assumptions (see the discussion in Appendix 1.C).

Note that bounds on many different functionals $f(P_U)$ can be computed *without modifying the set of constraints* in the program defined by (1.9); i.e. once the constraints for the model correspondence $\boldsymbol{G}$ have 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 $f(P_U)$ above, and then proving the sharpness of the derived bounds. As is shown in the application in Section 1.4, this feature allows the researcher to easily compute a variety of causal parameters to give a complete view of the effects of a program.

1.2.3 Identification with an Instrument

With an instrument $Z \in \mathcal{Z}$ we can consider the same model correspondence as before:

$$\boldsymbol{G}(y_0, y_1, \ldots, y_{K-1}) = \left\{ (y, d) \in \mathcal{W} \times \mathcal{Z} : y = \sum_{k=0}^{K-1} y_k \cdot \mathbb{1}\{d = k\} \right\}, \tag{1.10}$$

Here 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 $P_U \in \mathcal{P}_U^\dagger$ we must have $P_{U|Z}(A|Z = z) = P_U(A)$. For each $z \in \mathcal{Z}$, we can define the conditional capacity functional:

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

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

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

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

$$\mathcal{P}_U = \bigcap_{z \in \mathcal{Z}} \mathcal{P}_{U|Z}. \tag{1.11}$$

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

[10] Even if f does not meet these criteria, non-linear 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.

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

Chesher and Rosen (2017a), with an analogous result to that in Beresteanu et al. (2012) provided by their Theorem 4.

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

Example 2. *Consider again the case where $Y, D \in \{0,1\}$. Under the assumption $U \perp\!\!\!\perp Z$, the distribution of (Y_0, Y_1) is "unaffected" by the presence of the instrument in the sense that $P_{U|Z}(Y_0 \in A, Y_1 \in B|Z = z) = P_U(Y_0 \in A, Y_1 \in B)$. Since the outcome and treatment variable are binary we can still write the distribution of (Y_0, Y_1) as $P_U = (q_{00}, q_{10}, q_{01}, q_{11})$; however, we must now define the conditional distribution $P_{W|Z}(z) = (p_{00}(z), p_{10}(z), p_{01}(z), p_{11}(z))$, where $p_{ij}(z) = P_{W|Z}(Y = i, D = j|Z = z)$. Then Artstein's inequalities can be written for each fixed $Z = z$. Since Artstein's inequalities must hold for every $z \in \mathcal{Z}$, each inequality must hold for the infimum over the values of $z \in \mathcal{Z}$. For example, taking $A = \{(0,0),(0,1)\}$, Artstein's inequality for A when an instrument is available is given by:*

$$q_{00} + q_{01} \leq \inf_{z \in \mathcal{Z}} \{p_{00}(z) + p_{01}(z) + p_{11}(z)\}.$$

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 1.2.1 is then applicable to the case with an instrument where the identified set is as defined in equation (1.11); see Beresteanu et al. (2012) for additional discussion of this approach.

Mean independence can be accomplished in a similar manner, i.e. by writing all of Artstein's inequalities unconditional on Z and then imposing the constraint that $\mathbb{E}[Y_d] = \mathbb{E}[Y_d|Z = z]$ for all values of z. In addition, conditional joint independence $U \perp\!\!\!\perp Z|X$, where X is a vector of covariates, can also be easily accommodated by the method by writing the full set of Artstein's inequalities conditional $X = x$ for each x, and then intersecting over values of Z as above.

Note that there is no guarantee that the identified set defined by equation (1.11) is non-empty. By definition, emptiness of the identified set in (1.11) implies that there is 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 $\mathcal{P}_U^\dagger$. Thus, when $\mathcal{P}_U^\dagger$ 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).

1.2.4 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 1.2.1 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 1.2.1 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 1.2.1 is related to Proposition 1 in Torgovitsky (2016), which provides an analogous result but

for the class of complete econometric models.[12] Theorem 1.2.1 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 chapter. Since incomplete models nest complete models, Theorem 1.2.1 implies Proposition 1 in Torgovitsky (2016) when $\boldsymbol{G}$ 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 these 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 1.2.1 provides a bounding approach for discrete-valued outcomes, and requires solving only two optimization problems.

1.3 Computation and Estimation

Although Theorem 1.2.1 suggests a straightforward method of estimating bounds on functionals of the joint distribution, it may not always be computationally feasible. To appreciate the computational burden implied by Theorem 1.2.1, note that the identified set constructed via Artstein's inequalities is restricted by $2^{\min(|\mathcal{Y}|,|\mathcal{U}|)}-2$ constraints.[13] 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 follow the approach of Galichon and Henry (2011) and explore two methods of efficiently computing bounds on functionals of the joint distribution. The first method involves finding the smallest known collection of non-redundant constraints implied by Artstein's Theorem. The second method is based on reframing the bounding problem as an optimal transport problem. 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 decide between alternative approaches of computing the identified set. We then conclude the section by providing a result that shows estimation of the identified set under an optimization-based procedure is consistent.

1.3.1 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). In our context a core determining class is

[12] 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:

$$Y_{it}=Y_{it-1}U_{it}(1)+(1-Y_{it-1})U_{it}(0)=U_{it}(Y_{it-1}) \qquad \forall t\geq 1, \tag{1.12}$$

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 (1.12) 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 $\{Y_{it}\}_{t=0}^{T}$.

[13] 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 (1.5) or unobservables using the capacity functional given by (1.4).

any collection $\mathcal{S}$ of sets $A \in 2^{\mathcal{U}}$ such that if $P_U(A) \leq T(A)$ holds for all all $A \in \mathcal{S}$, then the same inequality holds for all $A \in 2^{\mathcal{U}}$.[14] This definition is consistent with the definition presented in Galichon and Henry (2011) and Chesher and Rosen (2017a). 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.

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.[15] 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 implied by Artstein's Theorem, even though 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 1.B for brevity.

By Lemma 1, we find that any set S in the exact core determining class is a collection of vectors $(y_0, \ldots, y_{K-1}) \in \mathcal{U}$ such that all vectors in S share exactly $K-1$ elements in common. Lemma 3 then says that if $K > 2$ *or* $|\mathcal{Y}| \leq K$, then the statement in the previous sentence completely characterizes the exact core determining class; otherwise, if $K = 2$ and $K < |\mathcal{Y}|$, then the exact core determining class contains precisely every set of at most $|\mathcal{Y}| - 1$ vectors with $K-1$ elements in common. Using these conditions, a researcher can easily select the *a priori* redundant constraints implied by Artstein's Theorem. Visual depictions of sets in the exact core class are given in Figure 1.2 in the case when $\mathcal{D} = \{0, 1\}$ to aid with interpretation of these conditions.

Using Lemmas 1, 2 and 3, it is also possible to show that in the POM with an instrument there are

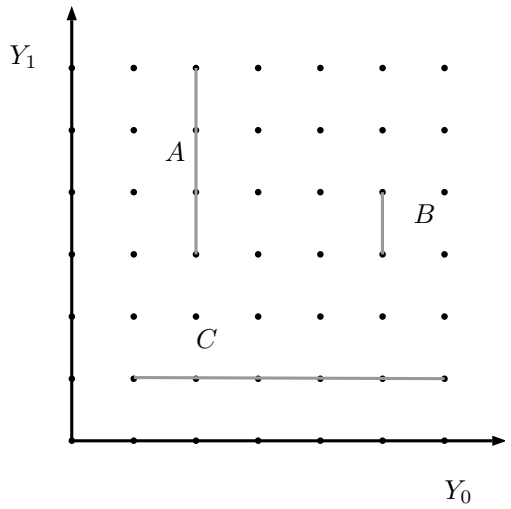
$$\begin{cases} \left(|\mathcal{Y}|^{|\mathcal{D}|} + \sum_{r=2}^{|\mathcal{Y}|} |\mathcal{Y}|^{|\mathcal{D}|-1} |\mathcal{D}| \binom{|\mathcal{Y}|}{r} - |\mathcal{Y}||\mathcal{D}| \right) \cdot |\mathcal{Z}|, & \text{if } |\mathcal{D}| = 2 \text{ and } |\mathcal{Y}| > |\mathcal{D}|, \\ \left(|\mathcal{Y}|^{|\mathcal{D}|}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

sets in the exact core determining class. The full version of this result is provided in Appendix 1.B. This result provides us with the number of sets in the exact core determining class for the POM, and helps us compare the exact core's computational tractability to other methods. Using our results, redundant inequalities can be efficiently identified and removed from the bounding problem.

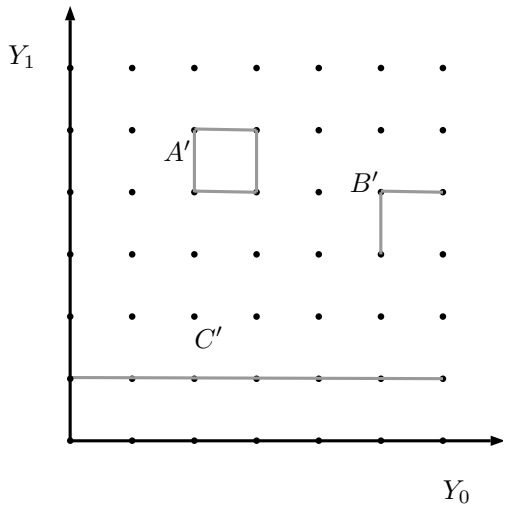
However, it is also important to note that under the exact core determining class approach of Luo and Wang (2016), it is not possible to compute the non-redundant constraints and then "intersect" by taking the infimum over all values of $z \in \mathcal{Z}$. Intuitively, this is because the results of Luo and Wang (2016) are valid only when the observed distribution P_W is a proper probability measure, and the infimum of $P_{W|Z}(\cdot|Z = z)$ over values of $z \in \mathcal{Z}$ is generally not a probability measure. Because of this feature, there are situations where the core determining class approach can be less computationally efficient then Artstein's inequalities; namely, when the support of the instrument is large relative to the support of both the observables (Y, D)

[14] For the case of a core determining class in a general incomplete model, we refer readers to Galichon and Henry (2011).

[15] 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.



(a) The sets A, B and C are examples of sets that are in the exact core determining class. In particular, A, B and C satisfy the conditions in Lemmas 1 and 3.



(b) The sets A', B' and C' are examples of sets that are NOT in the exact core determining class. In particular, A' and B' are redundant by Lemma 1, and C' is redundant by Lemma 3.

Figure 1.2: Examples of sets that are in, and are not in, the exact core determining class in the case of two potential outcomes (Y_0, Y_1).

and the support of the unobservable potential outcomes U.[16]

1.3.2 The Dual Approach

In this subsection we show how the duality result of Galichon and Henry (2006, 2011) can be used to construct an efficient estimation method.[17] To this end, let $\mathcal{M}_{\boldsymbol{G}|z}(P_{W|Z}, P_U)$ denote the set of Borel probability distributions, conditional on $Z = z$, with marginals $P_{W|Z}$ and P_U with support on $Graph(\boldsymbol{G})$ given by:

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

Then we can define the following set:

$$\mathcal{P}^*_{U|Z} = \{P_U \in \mathcal{P}^\dagger_U : \exists \pi \in \mathcal{M}_{\boldsymbol{G}|z}(P_{W|Z}, P_U)\}. \tag{1.13}$$

In other words, $\mathcal{P}_U$ defines the set of all distributions P_U such that there exists a joint distribution $\pi \in \mathcal{M}_{\boldsymbol{G}|z}(P_{W|Z}, P_U)$ that can rationalize the observed distribution $P_{W|Z}$ through the correspondence $\boldsymbol{G}$. In the empirical game setting of Galichon and Henry (2011), for example, this characterization is equivalent to establishing, for each candidate P_U, the existence of an equilibrium selection mechanism. Although the

[16] To see this, let $|\mathcal{S}_{ec}|$ represent the number of inequalities given in the exact core determining class, and let $|\mathcal{S}_a|$ represent the number of inequalities given by the full set of Artstein's inequalities. When an instrument is included we must write the inequalities in the exact core determining class for each value of $z \in \mathcal{Z}$, giving a total of $|\mathcal{Z}| \cdot |\mathcal{S}_{ec}|$ inequalities. However, when using the full set of Artstein's inequalities we can "intersect" over the values of $z \in \mathcal{Z}$. Thus, when using the full set of Artstein's inequalities with an instrument the number of inequalities needed is unchanged at $|\mathcal{S}_a|$. When the support of the instrument is large it is thus possible to have $|\mathcal{S}_a| < |\mathcal{Z}| \cdot |\mathcal{S}_{ec}|$.

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

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

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 P_U that satisfy Artstein's inequalities. The dual problem on the left is the one discussed in this subsection.

interpretation in our context is different, the idea in Galichon and Henry (2011) can still be applied. Define:

$$\mathcal{P}_U^* := \bigcap_{z \in \mathcal{Z}} \mathcal{P}_{U|Z}^*.$$

This collection is connected to the collection $\mathcal{P}_U$ characterized by Artstein's inequalities through Theorem 3 in Galichon and Henry (2011). Indeed, define:

$$\mathcal{P}_{U|Z} = \{P_U \in \mathcal{P}_U^\dagger : P_U(A) \leq T(A|Z=z) \text{ for all } A \in 2^{\mathcal{U}}\},$$

$$\mathcal{P}_{U|Z}^* = \{P_U \in \mathcal{P}_U^\dagger : \exists \pi \in \mathcal{M}_{\boldsymbol{G}|z}(P_{W|Z}, P_U)\}.$$

Then an application of Theorem 3 in Galichon and Henry (2011) shows that $\mathcal{P}_{U|Z} = \mathcal{P}_{U|Z}^*$, and thus $\mathcal{P}_U = \mathcal{P}_U^*$. From a practical perspective, this duality result provides an alternative method of estimation by showing that:

$$f^u := \sup_{P_U \in \mathcal{P}_U} f(P_U) = \sup_{P_U \in \mathcal{P}_U^*} f(P_U), \qquad f^\ell := \inf_{P_U \in \mathcal{P}_U} f(P_U) = \inf_{P_U \in \mathcal{P}_U^*} f(P_U). \tag{1.14}$$

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 $p_{ij}(z) = P_{W|Z}(Y=i, D=j|Z=z)$ and $q_{ij} = P_U(Y_0 = i, Y_1 = j)$. In addition, recall the set of non-redundant inequalities implied by Artstein's Theorem for this model:*

$$q_{00} \leq p_{00}(z) + p_{01}(z), \tag{1.15.1}$$
$$q_{01} \leq p_{00}(z) + p_{11}(z), \tag{1.15.2}$$
$$q_{10} \leq p_{10}(z) + p_{01}(z), \tag{1.15.3}$$
$$q_{11} \leq p_{10}(z) + p_{11}(z), \tag{1.15.4}$$
$$q_{00} + q_{01} \leq p_{00}(z) + p_{01}(z) + p_{11}(z), \tag{1.15.5}$$
$$q_{00} + q_{10} \leq p_{00}(z) + p_{01}(z) + p_{10}(z), \tag{1.15.6}$$
$$q_{01} + q_{11} \leq p_{00}(z) + p_{10}(z) + p_{11}(z), \tag{1.15.7}$$
$$q_{10} + q_{11} \leq p_{01}(z) + p_{10}(z) + p_{11}(z), \tag{1.15.8}$$

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 $\mathcal{P}_U$. Now let $\pi(z)$ be a vector with typical entry $\pi_{ij,k}(z) := P(Y_0(\omega) = i, Y_1(\omega) = j, D(\omega) = k|Z(\omega) = z)$, and consider the constraints:

$$\pi_{00,0}(z) + \pi_{01,0}(z) = p_{00}(z), \tag{1.16.1}$$
$$\pi_{00,1}(z) + \pi_{10,1}(z) = p_{01}(z), \tag{1.16.2}$$
$$\pi_{10,0}(z) + \pi_{11,0}(z) = p_{10}(z), \tag{1.16.3}$$
$$\pi_{01,1}(z) + \pi_{11,1}(z) = p_{11}(z), \tag{1.16.4}$$

where $\pi_{ij,k}(z) \geq 0$. Note that for any vector $\pi(z)$ satisfying (1.16.1)-(1.16.4), we can recover the vector $\mathbf{q}$ via:

$$\pi_{00,0}(z) + \pi_{00,1}(z) = q_{00}, \tag{1.17.1}$$

$$\pi_{01,0}(z) + \pi_{01,1}(z) = q_{01}, \tag{1.17.2}$$
$$\pi_{10,0}(z) + \pi_{10,1}(z) = q_{10}, \tag{1.17.3}$$
$$\pi_{11,0}(z) + \pi_{11,1}(z) = q_{11}. \tag{1.17.4}$$

In addition, 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* (1.16.1)-(1.16.4) *for* $z = 0$ *and* $z = 1$*, we must also impose the independence restriction given by:*

$$\pi_{00,0}(z) + \pi_{00,1}(z) = \pi_{00,0}(z') + \pi_{00,1}(z'), \tag{1.18.1}$$
$$\pi_{01,0}(z) + \pi_{01,1}(z) = \pi_{01,0}(z') + \pi_{01,1}(z'), \tag{1.18.2}$$
$$\pi_{10,0}(z) + \pi_{10,1}(z) = \pi_{10,0}(z') + \pi_{10,1}(z'), \tag{1.18.3}$$
$$\pi_{11,0}(z) + \pi_{11,1}(z) = \pi_{11,0}(z') + \pi_{11,1}(z'), \tag{1.18.4}$$

where $z, z' \in \{0, 1\}$*. By Theorem 3 in Galichon and Henry (2011), 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* (1.17.1) - (1.17.4) *from any probability vectors* $\pi(z)$*, for* $z = 0, 1$*, satisfying* (1.16.1) - (1.16.4) *and the independence restrictions* (1.18.1) - (1.18.4)*.*

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]

1.3.3 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 1.1. 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 1.2.

		$\|\mathcal{Y}\|$=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 1.1: 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 $D, Z \in \{0, 1\}$.

Table 1.1 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, the dual approach is hampered by the fact that it requires

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

		$\|\mathcal{Z}\|$=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 1.2: 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 $D, Y \in \{0, 1\}$.

a larger number of parameters which can introduce additional computational costs, especially when the objective function is non-linear.

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 generate the smallest number of constraints. Table 1.2 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. 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 $z \in \mathcal{Z}$. This property unique to Artstein's inequalities make them especially computationally tractable when the cardinality of Z is large. However, we note that even when $|\mathcal{Z}|$ is large, computation time using the dual approach can still be negligible if the functional f of interest is linear. However, for non-linear objective functions when the support $\mathcal{Z}$ is large and $|\mathcal{Y}|$ is small, using the characterization from Artstein's Theorem will allow for significant computational advantages, since the number of parameters (and thus the space over which we must optimize our functional of interest) will be much smaller.

Also note that neither Table 1.1 or 1.2 seem to support the use of the exact core approach, which is either dominated by the dual approach (in Table 1.1) or by Artstein's inequalities (in Table 1.2). 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.

These results also have implications for researchers who wish to perform inference on the resulting bounds. Recently there has been increased interest in inference problems for subvectors or functionals of a partially identified parameter vector; references include Chernozhukov et al. (2015), Bugni et al. (2017), Kaido et al. (2019a), Belloni et al. (2018), Gafarov (2019) and Cho and Russell (2019).[19] While a full discussion of the benefits and drawbacks of these inference procedures is beyond the scope of this chapter, we remark that it will be computationally easier to apply many of these inference procedures when there are less constraints, so that the results in this section are valuable in this regard as well. In addition, researchers should be aware that the number of constraints in the model—and whether the constraints are equality or inequality constraints—may also have implications for testing power, and thus may substantially affect the results of subvector and functional inference procedures. For a description of how to use the procedure in Chernozhukov et al. (2015) for a setting similar to the one in this chapter, we refer readers to the discussion in Torgovitsky (2016). However, we also remark that research on inference methods that exploit the full structure of the optimization-based bounds presented in this chapter is still an active area of research.

[19] While there have been many other papers in the literature on inference for the full partially identified vector—P_U in our case—we remark these procedures deliver confidence sets for subvectors or functionals that tend to be highly conservative (although still valid). See the introduction in Kaido et al. (2019a) for details of this issue of projection conservatism in partially identified models.

Overall, using a program that chooses the most computationally efficient 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 1.2.1, making the approach in this chapter tractable to run on a standard laptop computer for many bounding problems.

1.3.4 Estimation

To conclude this section, we show the conditions under which the optimization-based bounding procedure proposed in this chapter is consistent. Consistency is presented without an instrument for simplicity, but the result also holds when a instrument with finite support is available. Finally, the proof of consistency is given for the case when $\mathcal{P}_U$ is defined by Artstein's inequalities rather than the dual approach, although it is applicable to both approaches (since both approaches give numerically identical characterizations of $\mathcal{P}_U$). Proceeding, consider the usual empirical measure:

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

where $\{(Y_i, D_i)\}_{i=1}^n$ is an i.i.d. sample from the true distribution P_W. Let $\Theta^f(\mathbb{P}_n) = [f^\ell(\mathbb{P}_n), f^u(\mathbb{P}_n)]$ denote the estimated identified set for f, and let $\Theta^f(P_W) = [f^\ell(P_W), f^u(P_W)]$ denote the population identified set. Consistency in the estimation of sets is defined in terms of the Hausdorff distance d_H, which furnishes a metric on the space of non-empty compact subsets of $\mathbb{R}^d$.

Theorem 1.3.1. *Fix any continuous functional $f : \mathcal{P}_U \to \mathbb{R}$. Suppose that (a) $\mathcal{P}_U^\dagger$ is restricted only through linear (in)equality constraints; (b) the Jacobian of the linear equality constraints defining $\mathcal{P}_U$ (if any) has full row rank; (c) $\{W_i\}_{i=1}^n$ is i.i.d. from some probability measure P_W with finite support; and (d) int $(\mathcal{P}_U) \neq \emptyset$. Then $\Theta^f(\mathbb{P}_n) \xrightarrow{p} \Theta^f(P_W)$ in the Hausdorff metric.*

Note that Theorem 1.3.1 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. We refer the reader to Appendix 1.D for a full discussion of conditions required for the Theorem.

1.4 Application

We apply the results in this chapter 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. As discussed in Boozer and Cacciola (2001), attending small classes may raise student performance because of both a "pure classroom effect," arising from greater resources per student, as well as a peer-feedback (or "social multiplier") effect arising from long-term exposure to a higher quality peer environment. When decomposing the effects of Project STAR, Boozer and Cacciola (2001) find that the pure classroom effect accounts for most of the gains in student performance in the first year of exposure, but that the social-multiplier effect accounts for most of the gains in later years.

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 Boyd-Zaharias et al. (2007). As discussed in Krueger (1999), although the initial assignment of students to classrooms in Project STAR 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 chapter to provide various measures of the causal effects of the program on student performance. The outcome of interest is 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. Finally, consistent with the majority of studies on class size reduction policies, including Krueger (1999) and Krueger and Whitmore (2001), we will implicitly assume the stable unit treatment value assumption (SUTVA) of Rubin (1980); that is, $D_i = d$ implies $Y_i = Y_{di}$ (c.f. Angrist et al. (1996)).[20] Violations of this assumption could occur, for example, if the researcher believes that students' potential outcomes should be defined with respect to the exact number of students in the classroom, rather than the dichotomous "small" and "large" distinction, or that a student's potential outcomes should be defined with respect to the specific identity of his/her classmates. However, for simplicity of exposition we will abstract away from these concerns and will leave the exploration of these issues using our method for future research.

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 1.3. 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

[20] Note that using the percentile *ranks* as an outcome variable mechanically violates SUTVA in finite sample, although the mechanical violation becomes negligible in large samples. I thank an anonymous referee for pointing this out.

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 is exactly 50/100, so that the values in the table can be interpreted relative to this number. Notice in Table 1.3 that sample characteristics are well-balanced across $Z = 1$ and $Z = 0$, which provides some evidence that the randomization was successful.[21] 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.

Table 1.3: 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 Math 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.

For the purpose of the application, percentile ranks for math and reading scores were averaged together to create a single outcome variable, as in Krueger (1999). Also as in Krueger (1999), the treatment effects studied in the application will be "reduced-form" in the sense that they will aggregate the pure classroom effect and social multiplier effect studied in Boozer and Cacciola (2001), and discussed at the beginning of this section, into a single treatment effect for small class sizes. 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 $P(Y_1 > Y_0) \geq 0.95$. This implies that we consider only data-generating processes under which students strictly benefit from small classes sizes with 95% probability. This corresponds with our prior belief that smaller class sizes are beneficial to most students, consistent with the evidence in Krueger (1999) and Krueger and Whitmore (2001) on the impact of the Tennessee STAR program.

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

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). It is also important to keep in mind 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.

(i) $P(Y_0 \leq Median(Y), Y_1 > Median(Y))$*: 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*. Since Median(Y) is the observed outcome, and not necessarily the median of the counterfactual outcomes* Y_0 *and* Y_1*, 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 as a result of attending a small versus regular classroom.*

(iii) $P(Y_1 > Y_0)$*: The voting criterion, which measures the proportion of students whose rank strictly improves from attending smaller classrooms.*

(iv) $P(Y_1 > Median(Y) | Y_0 \leq Median(Y))$*: 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.*

(v) $P(Y_0 \leq Median(Y))$*: The proportion of people who would have a below median rank in the regular classroom. Therefore, if* $P(Y_0 \leq Median(Y)) > 0.5$*, for example, then we know that the median of the unobserved rank* Y_0 *is less than the median of* Y*. Using this method we could also recover information on other quantiles.*

(vi) $P(Y_1 > Median(Y))$*: The proportion of people who would have an above-median rank in the small classroom. See the discussion in the previous point.*

(vii) $Corr(Y_0, Y_1)$*: The correlation between student ranks in regular versus small classrooms. A positive correlation indicates that students with low ranks in regular class rooms are also likely to have low ranks in small classrooms.*

(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.*

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 1.4. First note that Table 1.4 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 1.4, the joint probability $P(Y_0 \leq Median(Y), Y_1 > Median(Y))$ 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 $[0.04, 0.22]$. An extended analysis for this parameter is given in Appendix 1.E.

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

Table 1.4: Bounds on School Achievement from the Tennessee STAR Experiment Assuming $P(Y_1 > Y_0) \geq 0.95$.**

		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
$P(Y_0 \leq Median(Y), Y_1 > Median(Y))$:†	Bins=25	0.09	0.26	0.04	0.22
	Bins=30	0.08	0.26	0.04	0.22
	Bins=35	0.09	0.26	0.04	0.22
$\mathbb{E}[Y_1 - Y_0]$:	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
$P(Y_1 > Y_0)$:*	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
$P(Y_1 > Median(Y) \mid Y_0 \leq Median(Y))$:†	Bins=25	0.14	0.97	0.07	1.00
	Bins=30	0.14	0.96	0.07	1.00
	Bins=35	0.17	1.00	0.07	1.00
$P(Y_0 \leq Median(Y))$:†	Bins=25	0.55	0.55	0.52	0.52
	Bins=30	0.55	0.55	0.52	0.52
	Bins=35	0.53	0.53	0.52	0.52
$P(Y_1 > Median(Y))$:†	Bins=25	0.52	0.66	0.50	0.66
	Bins=30	0.50	0.66	0.49	0.65
	Bins=35	0.53	0.69	0.48	0.65
$Corr(Y_0, Y_1)$:	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

†: Recall that Median(Y) is the median of the observed outcome, but not necessarily the median of Y_0 or Y_1.
∗: The parameter $P(Y_1 > Y_0)$ is the only parameter estimated without the MTR assumption $P(Y_1 > Y_0) \geq 0.95$.
∗∗: All values are computed using the "plug-in" estimator for the empirical probabilities, as described in Appendix 1.D.

from attending small class sizes, and are consistent with the results of Krueger (1999) and Krueger and Whitmore (2001).[22]

The voting criterion $P(Y_1 > Y_0)$ is the only parameter estimated without imposing the MTR assumption.[23] 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 $[0.05, 0.98]$. This provides evidence that the MTR assumption restricting $P(Y_1 > Y_0) \geq 0.95$ may have substantial identifying power when estimating the other parameters in Table 1.4. This is explored further in Appendix 1.E.

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 $P(Y_1 > Median(Y) \mid Y_0 \leq Median(Y))$ range from $[0.17, 1.00]$ and for the grade 8 outcomes it ranges from $[0.07, 1.00]$. An extended analysis for this parameter is also provided in Appendix 1.E.

Bounds on the marginal probabilities $P(Y_0 \leq Median(Y))$ and $P(Y_1 > Median(Y))$ are found to be informative, taking values respectively in the intervals $[0.53, 0.53]$ and $[0.53, 0.69]$ for the grade 3 outcomes, and $[0.52, 0.52]$ and $[0.48, 0.65]$ for the grade 8 outcomes. In particular, note that the value of $P(Y_0 \leq$

[22] 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.

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

$Median(Y)$) is nearly point-identified in this application. Upon further investigation this is found to be a result of the fact the estimated value of $P(D = 1|Z = 0)$ is nearly zero in our sample.[24]

Bounds on the correlation coefficient are found to be marginally informative, ranging in $[0.02, 0.5]$ 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. However, sensitivity analysis in Appendix 1.E shows identification power for this parameter is likely coming from our MTR assumption $P(Y_1 > Y_0) \geq 0.95$.

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 $P(Y_1 > Y_0) \geq 0.95$, Appendix 1.E contains bounding results when this assumption is relaxed to $P(Y_1 > Y_0) \geq 0.5$. 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 chapter (which always delivers 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 chapter can be used to identify bounds on 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.

1.5 Conclusion

This chapter 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 a 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 modelling assumptions, such as the monotone treatment response, and monotone instrumental variables assumptions (see the discussion in Appendix 1.C). Finally, we show an application of the results to the Tennessee STAR experimental data.

[24] For intuition, note that the value of $P(Y_0 \leq y)$ can be decomposed as:

$$\begin{aligned} P(Y_0 \leq y) &= P(Y_0 \leq y|Z = 0) \\ &= P(Y \leq y|D = 0, Z = 0)P(D = 0|Z = 0) + P(Y_0 \leq y|D = 1, Z = 0)P(D = 1|Z = 0), \end{aligned}$$

where the first equality follows from independence. Partial identification of $P(Y_0 \leq y)$ results from lack of knowledge of $P(Y_0 \leq y|D = 1, Z = 0)$. Ignoring the MTR assumption, when $P(D = 1|Z = 0) \approx 0$ the second (unknown) term in the previous display is negligible, implying the identified set for $P(Y_0 \leq y)$ has a small length.

Appendix 1.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 $\mathcal{X}$.[25] Fix some probability space $(\Omega, \mathfrak{A}, P)$, and let $\mathbf{X} : \Omega \to \mathcal{F}$.

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

$$\{\omega : \mathbf{X}(\omega) \cap A \neq \emptyset\} \in \mathfrak{A}.$$

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

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

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

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

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

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

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:*

$$\mu(X \in A) \leq \nu(\mathbf{X} \cap A \neq \emptyset), \qquad \forall A \in \mathcal{K}. \tag{1.19}$$

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

Appendix 1.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 1.B.1 (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*

$$P_U^*(A) > P_W(\boldsymbol{G}^{-1}(Y,D) \cap A \neq \emptyset),$$

where

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

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

Definition 1.B.2 (Luo and Wang (2016)). *Let $\mathcal{S}_u$, $\mathcal{S}_w$ and $\mathcal{S}_w^{-1}$ be the collections of sets with the following properties:*

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

(i) A is self-connected.[26]

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

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

(i) B is self-connected.

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

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

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

Theorem (Luo and Wang (2016)). *Assume that the bipartite graph represented by $\mathcal{G} = (\mathcal{W}, \mathcal{U}, \boldsymbol{G})$ is connected; that is, for every $A_1, A_2 \subset \mathcal{U}$ such that $A_1, A_2 \neq \emptyset$ and $A_1 \cup A_2 = \mathcal{U}$ we have $\boldsymbol{G}(A_1) \cap \boldsymbol{G}(A_2) \neq \emptyset$. If the measure P_W on $\mathcal{W}$ is non-degenerate, i.e. $P_W(W = w)$ is non-zero for all $w \in \mathcal{W}$, then the exact core determining class is given by:*

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

[26] 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 $\boldsymbol{G}(A_1) \cap \boldsymbol{G}(A_2) \neq \emptyset$.

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.[27] 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 $\boldsymbol{G}$. The structure of the correspondence $\boldsymbol{G}$ in the POM is best illustrated when looking at the bipartite graph $\mathcal{G} = (\mathcal{W}, \mathcal{U}, \boldsymbol{G})$. 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 $\mathcal{W}$ of the graph $\mathcal{G}$ has exactly $|\mathcal{Y}||\mathcal{D}|$ nodes with degree $|\mathcal{Y}|^{|\mathcal{D}|-1}$.

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

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

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

Lemma 1.B.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 1.B.2. *For the POM we have*

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

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

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

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

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

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

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

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

Lemma 1.B.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, $\mathcal{S}_u \subset \mathcal{S}_w^{-1}$.*

[27] 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.

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

Theorem 1.B.1. *Suppose that the distribution P_W 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 1.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 $q_{ij} = P(Y_0 = i, Y_1 = j)$, and suppose we wish to bound the parameter

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

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

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

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

$$\begin{aligned} q_{00} &= \pi_{00,0} + \pi_{00,1}, \\ q_{01} &= \pi_{01,0} + \pi_{01,1}, \\ q_{10} &= \pi_{10,0} + \pi_{10,1}, \\ q_{11} &= \pi_{11,0} + \pi_{11,1}. \end{aligned}$$

Then the optimization problem is:

$$\max_{\pi} \frac{\pi_{01,0} + \pi_{01,1}}{\pi_{00,0} + \pi_{00,1} + \pi_{01,0} + \pi_{01,1}}, \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \pi = \mathbf{p}, \\ \mathbf{0} \preccurlyeq \pi \preccurlyeq \mathbf{1}. \end{cases} \tag{1.20}$$

To write this as a linear programming problem, define

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

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

Then the problem above can be re-written as:

$$\max_{\tilde{\pi}, r} c' \cdot \tilde{\pi}, \qquad s.t. \qquad \begin{cases} \mathbf{A}_\pi \cdot \tilde{\pi} - \mathbf{p} \cdot r = 0, \\ d_1 \cdot \tilde{\pi} = 1, \\ d_2 \cdot \tilde{\pi} - r = 0, \\ \mathbf{0} \preccurlyeq \tilde{\pi} \preccurlyeq \mathbf{1}, \\ r \geq 1. \end{cases} \tag{1.21}$$

This can be seen by replacing the objective function in (1.20) with the equivalent objective function in (1.21), by multiplying both sides of the constraint $\mathbf{A}_\pi \cdot \pi = \mathbf{p}$ in (1.20) by the variable r and rearranging, and by

imposing constraints ensuring that the conditional probability measure is a proper probability measure, namely:

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

Alternatively, we could write the same problem more compactly as

$$\max_{\tilde{\mathbf{q}}_r} c_r' \cdot \tilde{\mathbf{q}}_r \qquad s.t. \qquad \begin{cases} \mathbf{A}_r \cdot \tilde{\mathbf{q}}_r = \mathbf{a}_r, \\ b_l \preccurlyeq \tilde{\mathbf{q}}_r \preccurlyeq b_u, \end{cases} \tag{1.22}$$

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

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

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

The problem (1.22) 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.

1.C.1 Introducing Additional Constraints

Imposing additional assumptions on the unobserved probability measure P_U in an analytic framework requires a new proposed identified set and corresponding proof of sharpness. In contrast, additional assumptions can be imposed easily on P_U in the computational framework. In addition, in many cases additional assumptions can be included as linear constraints in P_U, 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 P_U 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 1.C.1 (MTR, Manski and Pepper (2000))**.** *Let $\mathcal{Y}_d$ be an ordered set. Then the MTR assumption is satisfied if* $d' \geq d \implies P(Y_{d'} \geq Y_d) = 1$.

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 1.C.2 (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 P_U. 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 P_U (which makes them especially amenable to inclusion in linear programs). Since the set $\mathcal{P}_U^\dagger$ is still convex and closed under these constraints, estimation using Artstein's inequalities is consistent by Theorem 2. The MTR and MIV assumptions presented are examples of additional assumptions that can be imposed to obtain a more informative analysis, although there are many other assumptions that might also be imposed without affecting any of the previous results.

Appendix 1.D Consistency and Inference

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

Consider the usual empirical measure:

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

where $\{(Y_i, D_i)\}_{i=1}^n$ is an i.i.d. sample from P_W. Now define the set:

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

or equivalently:

$$\mathcal{P}_U(\mathbb{P}_n) := \{P_U \in \mathcal{P}_U^\dagger : \exists \pi \in \mathcal{M}_{\boldsymbol{G}}(\mathbb{P}_n, P_U)\}.$$

Consistency in the estimation of sets is usually defined in terms of the Hausdorff distance d_H, which furnishes a metric on the space of non-empty compact subsets of $\mathbb{R}^d$.[28] Here we are interested in establishing consistency with respect to the Hausdorff metric of the set:

$$\Theta^f(\mathbb{P}_n) = [f^\ell(\mathbb{P}_n), f^u(\mathbb{P}_n)] \quad \text{with} \quad f^\ell(\mathbb{P}_n) = \sup_{P_U \in \mathcal{P}_U(\mathbb{P}_n)} f(P_U), \quad f^u(\mathbb{P}_n) = \inf_{P_U \in \mathcal{P}_U(\mathbb{P}_n)} f(P_U), \tag{1.23}$$

for the set:

$$\Theta^f(P_W) = [f^\ell(P_W), f^u(P_W)] \quad \text{with} \quad f^\ell(P_W) = \sup_{P_U \in \mathcal{P}_U} f(P_U), \quad f^u(P_W) = \inf_{P_U \in \mathcal{P}_U} f(P_U). \tag{1.24}$$

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

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

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

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

Condition (d) is worth some discussion. Note that Theorem 2 shows that estimation of bounds on any continuous functional of the joint distribution can be completed using Artstein's inequalities without the

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

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

need for a tuning parameter. However, this is done at the cost of ruling out point identification through assumption (d). While point identification is a knife-edge case under all assumptions considered in this chapter, some researchers may feel assumption (d) is too restrictive. If this is the case, researchers can add a slackness term that drifts towards zero—say c_n—to each of the inequalities defining the set $\mathcal{P}_U$, and Theorem 2 can then be applied with assumption (d) replaced with the assumption that $\mathcal{P}_U \neq \emptyset$. A general rule for selecting the slackness is that it should dominate relative to sampling error; thus, a possible choice for the slackness is given by $c_n = \sqrt{\log(n)/n}$. Introducing such a slackness term will cause any estimated identified sets to have slightly larger length, although any difference will be negligible for large n.

Appendix 1.E Application Robustness Exercise

Figure 1.3 shows plots of $P(Y_1 > y_q | Y_0 \leq y_{0.5})$ and $P(Y_1 > y_{0.5} | Y_0 \leq y_q)$ 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 1.4 shows informative plots of the joint distribution $P(Y_1 > y_q, Y_0 \leq y_{0.5})$ and $P(Y_1 > y_{0.5}, Y_0 \leq y_q)$ against y_q.

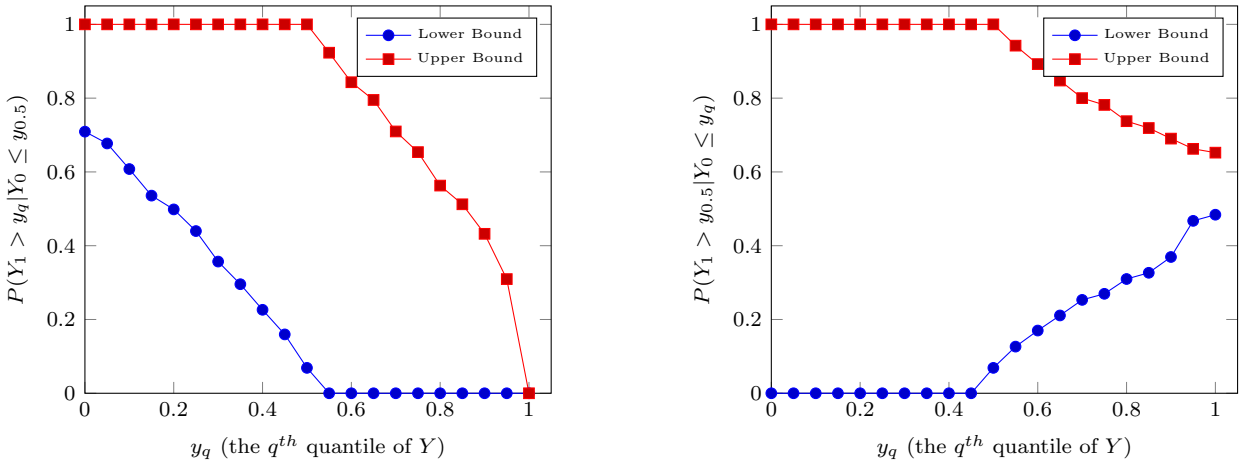


Figure 1.3: Bounds on the conditional probability (Grade 3, Bins=35, MTR assumption $P(Y_1 > Y_0) \geq 0.95$).

Table 1.5 shows bounds for the parameters of interest in the Tennessee STAR experiment when the MTR condition is relaxed from $P(Y_1 > Y_0) \geq 0.95$ to the MTR condition $P(Y_1 > Y_0) \geq 0.5$. As discussed in the main text, the bounds on some of the parameters—such as the bounds on $P(Y_1 > Median | Y_0 \leq Median)$, $P(Y_0 \leq Median)$, $\sqrt{Var(Y_0)}$)—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 condi-

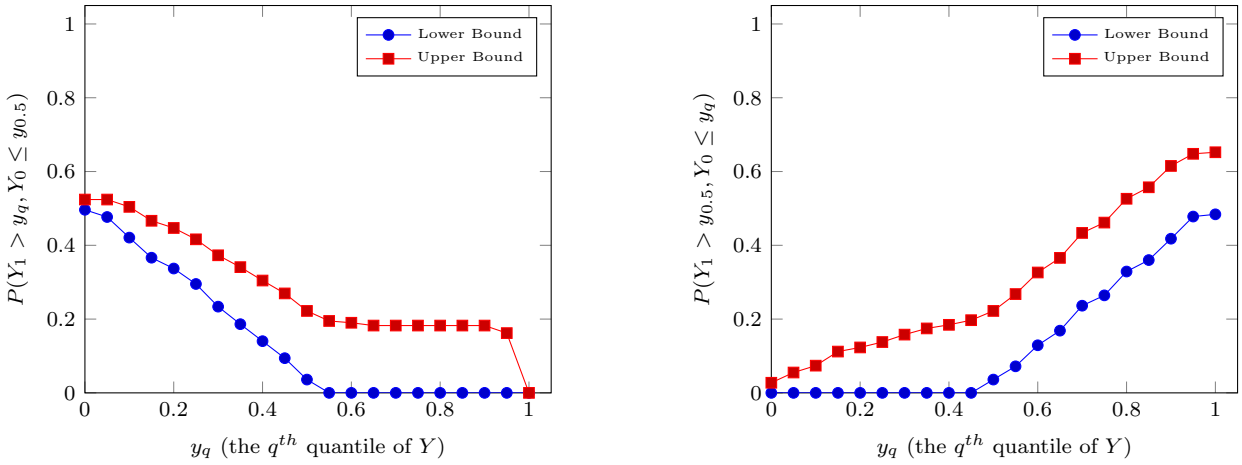


Figure 1.4: Bounds on the joint probability (Grade 3, Bins=35, MTR assumption $P(Y_1 > Y_0) \geq 0.95$).

tion ($P(Y_1 > Y_0) \geq 0.95$ or $P(Y_1 > Y_0) \geq 0.5$) 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 chapter (which always delivers sharp bounds).

Table 1.5: Bounds on School Achievement from the Tennessee STAR Experiment Assuming $P(Y_1 > Y_0) \geq 0.5$

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

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

Appendix 1.F Proofs

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

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

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

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

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

Now consider the forward direction; i.e. suppose that $A \in \mathcal{S}_u$ and $|A| = r \geq 2$, and proceed by inducting on r. First consider the case when $r = 2$. For any $A \in \mathcal{S}_u$ with $|A| = 2$, take the singletons u_1, u_2 that comprise A (i.e. the singletons such that $u_1 \cup u_2 = A$). If u_1 and u_2 share more than $K-1$ elements then they are the same vector. It is also clear that u_1 and u_2 must share at least one element, otherwise condition (a)(i) in Definition 1.B.2 is not satisfied. Thus, suppose u_1 and u_2 share $1 \leq k < K-1$ elements. Without loss of generality, suppose that they share the first k elements, so that we can write the vectors u_1 and u_2 as:

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

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

Now consider the vector u_3 given by:

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

I.e. u_3 is the vector that shares the same first k elements with both u_1 and u_2, shares the next $(K-1)-(k+1)$ elements with vector u_1, and shares the last element with vector u_2. Clearly this vector u_3 exists, $u_3 \notin A$ and $\boldsymbol{G}(u) \subset \boldsymbol{G}(u_1 \cup u_2)$, 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 $u \in \mathcal{U}$. 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 1.B.2). Suppose without loss of generality that this vector is $u_i = u_1$; this simplification is only to reduce the level of abstraction. Now consider two cases:

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

2. $u_{\ell+1}$ and u_1 share at least one of the elements $y_0, y_1, \ldots, y_{K-1}$: Note that if these elements share y_{1K} then we are in the previous case, since this implies that they do not share at least one element in

$y_0, y_1, \ldots, y_{K-1}$. Thus, suppose they do not share y_{1K}. If they share all other elements, then $u_{\ell+1}$ shares exactly $K-1$ elements with all vectors in A, which is a contradiction. Thus, there must exist at least one element in $y_0, y_1, \ldots, y_{K-1}$ that they do not share. But note there exists a $u \in \mathcal{U}$ that is the same as u_1 except that its last element is replaced with the last element of $u_{\ell+1}$. But then $u \notin A'$ and $\boldsymbol{G}(u) \subset \boldsymbol{G}(A')$, 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 1.B.2. For notational simplicity, let $M := |\mathcal{Y}|$ and $K := |\mathcal{D}|$.

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

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

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

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

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

■

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

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

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

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

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

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

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

Now set

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

and consider the follow cases:

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

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

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

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

so $B \not\subseteq \mathcal{W}_k$ for any k and $\mathcal{W}_k \not\subseteq B$ for any k, which proves $B \in \mathcal{S}_w$ by Lemma 1.B.2.

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

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

contains elements from $\mathcal{W}_2, \ldots, \mathcal{W}_k$ regardless of the magnitude of r, and $\mathcal{W}_k \not\subseteq B$ for any k. Thus by Lemma 1.B.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 $A^c = \boldsymbol{G}^{-1}(B)$, 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 $A^c = \boldsymbol{G}^{-1}(B)$, so that $A \in \mathcal{S}_w^{-1}$. This completes the proof. ■

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

1. Note that every singleton trivially satisfies the conditions in Definition 1.B.2, 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.B.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 1.B.2, to construct a set $B \in \mathcal{S}_w$ of size r from the singletons we can choose r elements from any combination of the K subsets $\mathcal{W}_k$, but we must choose elements from at least two subsets, and we must choose less than M elements from each collection. Now note that there are $\binom{K}{\ell}$

ways to choose from any $2 \leq \ell \leq K$ collections, and $\binom{M}{v_k}$ ways to choose $1 \leq v_k \leq M-1$ elements from each collection. Finally, we must ensure that if we are constructing an r-element set B that we have

$$\sum_k v_k = r.$$

Combining everything, there are

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

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

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

as claimed.

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

■

Proof of Theorem 2. Notation for the proof is given in Appendix 1.D.

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

$$f^u(\mathbb{P}_n) = \sup_{Q \in \mathcal{P}_U(\mathbb{P}_n)} f(P_U). \tag{1.25}$$

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

$$\limsup_{n \to \infty} P\left(|f^u(\mathbb{P}_n) - f^u(P_W)| > \varepsilon \right) = 0. \tag{1.26}$$

Now note:

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

Let Δ_U denote the $(|\mathcal{U}|-1)$-simplex. Since $\Delta_U \subset \mathbb{R}^{d_U}$ is compact, continuity of f implies uniformly continuity over Δ_U. Thus, we know that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $||P_U - P_U'|| < \delta$

implies $|f(P_U) - f(P_U')| < \varepsilon$. Thus, to show (1.26) it suffices to show that for every $\delta > 0$:

$$\limsup_{n\to\infty} P\left(d_H(\mathcal{P}_U(\mathbb{P}_n), \mathcal{P}_U) > \delta\right) = 0. \tag{1.27}$$

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

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

$$\implies \lambda_k = 1 - \sum_{j\in I(\mathcal{U})} P_U(U = u_j) - \sum_{j\in I(\mathcal{U}), j\neq k} \lambda_j, \tag{1.28}$$

and then add the non-negativity constraint on (1.28) (where $I(\mathcal{U})$ is an index set for elements in $\mathcal{U}$). Thus, there will be $(d_e - 1)$ equality constraints in the vector $g(\theta, P_W)$, and d_θ inequality constraints given by the vector:

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

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

$$T(\theta, P_W) = g(\theta, P_W)' g(\theta, P_W).$$

Then under assumption (d) we have:

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

The sample analog of the above is:

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

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

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

2. The $(|\mathcal{W}| - 1)$-simplex $\Delta_W \subset \mathbb{R}^{d_{\mathcal{W}}}$ is compact. Θ is also compact (since it is without loss of generality that we restrict $\lambda \in [0, 1]$).

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

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

Consistency of $\Theta(\mathbb{P}_n)$ for $\Theta(P_W)$ in the Hausdorff metric then follows from Theorem 2.1 in Shi and Shum (2015). This in turn implies consistency of $\mathcal{P}_U(\mathbb{P}_n)$ for $\mathcal{P}_U$ in the Hausdorff metric, and thus completes the proof.

■

Chapter 2

Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments

This chapter considers uniformly valid (over a class of data generating processes) inference for linear functionals of partially identified parameters in cases where the identified set is defined by linear (in the parameter) moment inequalities. We propose a bootstrap procedure for constructing uniformly valid confidence sets for a linear functional of a partially identified parameter. The proposed method amounts to bootstrapping the value functions of a linear optimization problem, and subsumes subvector inference as a special case. In other words, this chapter shows the conditions under which "naively" bootstrapping a linear program can be used to construct a confidence set with uniform correct coverage for a partially identified linear functional. Unlike other proposed subvector inference procedures, our procedure does not require the researcher to repeatedly invert a hypothesis test, and is extremely computationally efficient. In addition to the new procedure, the paper also discusses connections between the literature on optimization and the literature on subvector inference in partially identified models. This chapter was written jointly with my peer and friend, JoonHwan Cho.

2.1 Introduction

This chapter proposes a uniformly valid (over a large class of data generating processes) inference procedure for a linear functional ψ of a partially identified parameter vector θ in models with linear (in θ) moment functions. In particular, the paper proposes to use a "naive" bootstrap procedure to approximate the distribution of the endpoints of the projected identified set, and discusses conditions under which the procedure is uniformly valid.

The main idea is to use results from the Operations Research literature that allow the researcher to approximate the distribution of the value functions in linear programs with stochastic constraints using a functional delta method. The contribution of this chapter is to use these results for stochastic programs as a proof device to show the uniform validity of a simple bootstrap procedure for constructing confidence sets for subvectors or functionals of the identified set in partially identified econometric models. Intuitively, bounding a linear functional over an identified set defined by linear moment functions amounts to solving two linear optimization problems: one maximization problem for the upper bound, and one minimization problem for the lower bound. Thus, the endpoints of the identified set for a functional of a partially-identified

parameter can be viewed as *value functions* of two "stochastic programs." An inference procedure is then constructed by decomposing perturbations in the optimal value function into perturbations arising from the objective function and perturbations arising from the constraint functions. By the envelope theorem, perturbations in the constraints are related to the value functions through the Lagrange multipliers. The total effect of perturbations in the constraints on the value function is then given by a weighted sum of the perturbations in all binding constraints, where the weights are determined by the Lagrange multipliers. Through this mechanism, we can relate the distribution of the binding moment functions to the distribution of the value function of a stochastic program.

To prove the validity of our procedure requires noticing that, under some conditions, the value functions in linear stochastic programs are *Hadamard directionally differentiable* with respect to perturbations in the underlying probability measure. However, this form of differentiability is not sufficient for *uniformly* valid confidence sets. This result relates to Kasy (2019), who emphasizes that failures of uniformity often result as failures of the uniform versions of the delta method. We demonstrate the conditions under which the value functions of a linear stochastic program satisfy the natural definition of *uniform* Hadamard directional differentiability with respect to perturbations in the underlying probability measure. The condition that emerges as being most important for our procedure is the existence and uniqueness of optimal solutions and Lagrange multipliers.

Uniform Hadamard directional differentiability is sufficient for us to prove the validity of a simple uniformly valid bootstrap procedure to estimate the confidence set for a functional of interest. In the environment considered in this chapter, bounds on the linear functional of interest can always be constructed by solving linear programs, and our bootstrap procedure amounts to repeatedly solving analogous "bootstrap linear programs." Given it's simplicity, we call the process of repeatedly solving these bootstrap linear programs the "naive" bootstrap approach to functional inference in partially identified models. Following this approach, a confidence set for the partially identified functional of interest is constructed by selecting appropriate quantiles from these value function bootstrap distributions. In other words, this chapter shows the conditions under which naively bootstrapping a linear program can be used to construct a confidence set with uniform correct coverage for a partially identified linear functional.

This naive bootstrap approach has considerable advantages relative to other approaches. In particular, it does not require repeatedly inverting a hypothesis test, and thus is very computationally efficient—also owing to the computational efficacy of linear programming—and promising for cases when the parameter vector of interest is high-dimensional. Indeed, in our main simulation exercise we find it takes only about 10 seconds to compute a two-sided confidence set for a functional of a partially identified parameter vector with 400 elements. Interestingly, the use of Lagrange multipliers allows us to avoid rescaling the moment conditions by their sample standard deviations in our procedure. This is in contrast to other comparable methods. Intuitively, any rescaling of the moment functions is countered by an equivalent (but opposite) rescaling of the Lagrange multipliers. This ensures our bootstrap procedure remains a sequence of linear—and thus easy to solve—optimization problems. Furthermore, the assumption of a uniform constraint qualification turns out to be sufficient to allow the user to avoid using moment selection procedures (see Andrews and Soares (2010)), which are common in the literature on partial identification. One of our contributions will be to highlight some interesting comparisons and contrasts between assumptions from the optimization literature versus the assumptions from the previous literature in partial identification.[1]

Subvector inference, or inference on functionals of the identified set, has recently been a topic of considerable interest in the partial identification literature. The earlier papers of Andrews and Guggenberger (2009) and Andrews and Soares (2010) propose to project confidence sets constructed for the entire parameter vector in order to obtain confidence sets for a particular subvector of interest. While these procedures are

[1] This is also done in a recent paper by Kaido et al. (2019b).

uniformly valid, they can be highly conservative when the dimension of the partially identified parameter vector is large (see the discussion in Kaido et al. (2019a)). Both Romano and Shaikh (2008) and Bugni et al. (2017) consider inverting profiled test statistics in order to construct confidence sets for subvectors or functionals, where Romano and Shaikh (2008) construct critical values using subsampling and where Bugni et al. (2017) derive the asymptotic distribution for their profile test statistic for a large class of test functions. Bugni et al. (2017) show that their test dominates projection-based procedures in terms of asymptotic power, and they derive conditions under which it dominates the subsampling-based approach of Romano and Shaikh (2008). Kaido et al. (2019a) provide a "calibrated projection" inference method for functionals of a partially identified parameter. Intuitively, this procedure suitably relaxes the model's moment inequalities, and then solves two optimization problems subject to the relaxed constraints in order to obtain the endpoints of the confidence interval for the functional of interest. The relaxation of the constraints requires the correct calibration of a relaxation parameter in order to obtain uniformly correct coverage. Kaido et al. (2019a) first linearize any nonlinear moment functions, and then propose an efficient algorithm to calibrate the relaxation parameter. This allows their procedure to be computationally attractive relative to other methods in nonlinear models. Similar to the method proposed here, the method of Kaido et al. (2019a) does not invert a test statistic.

The overall approach to constructing confidence sets in this chapter is most closely related to the approach in Gafarov (2019), who also shows how to construct uniformly valid confidence sets for linear functionals of a partially identified parameter in an optimization framework. It is well known from Hirano and Porter (2012) that it is impossible to obtain a locally unbiased estimator of the value function when the value function is nondifferentiable, and to address these problems Gafarov (2019) proposes including a regularization term in the objective function to ensure a unique optimal solution is selected. In contrast we assume the existence of unique optimal solutions. Similar to Gafarov (2019), we also impose a linear independence constraint qualification to ensure uniqueness of the Lagrange multipliers. However, we allow for the linear functional in the bounding problem to be data-dependent, and both our bootstrap procedure and our proof of uniform validity are very different. Overall, we believe our contribution is both practical and theoretical, and complements this recent work by Gafarov (2019).

The main proofs of this chapter uses results from Shapiro et al. (2009). The main result used from Shapiro et al. (2009) is the proof of Hadamard directional differentiability of value functions for stochastic programs. However, we extend this result by showing the conditions under which the value functions for a stochastic program satisfy uniform Hadamard directional differentiability, which is sufficient to derive a uniform delta method result.

Throughout the paper we use notation standard in empirical process theory; in particular, the expectation of a random element X_t with respect to a measure P is given by PX_t. If the random element X_t is a vector, then the expectation is interpreted element-wise. The random variables $W_1, W_2, \ldots, W_n$ are assumed to be coordinate projections from the product space $(\mathcal{W}^n, \mathcal{A}^n, P^n)$, where $P^n = P \otimes P \otimes \ldots \otimes P$, and we will denote $(\mathcal{W}^\infty, \mathcal{A}^\infty, P^\infty)$ as the infinite product space. The empirical measure is represented by $\mathbb{P}_n$, which is implicitly a function of the generating measure P_n at sample size n. We index estimated quantities by the empirical distribution; for example, rather than $\hat{\theta}$, we write $\theta(\mathbb{P}_n)$. This is done to emphasize the underlying measure relevant to the construction of the parameter, and becomes useful in both the discussion and the proofs of the main results. Finally, we use $||\cdot||$ to denote the euclidean norm throughout. For the most part, we will avoid issues of measurability as much as possible, although all the proofs follow from the definition of weak convergence in the sense of Hoffmann-Jørgensen (1991) (c.f. Van Der Vaart and Wellner (1996) Chapters 1.1 and 1.2).

2.2 Overview of Results and Motivating Examples

2.2.1 Main Ideas

This subsection will discuss simplified versions of the main ideas in the paper before the technical details are introduced in the next section. Our main motivation is to construct uniformly valid confidence sets for the expectation of the random objective function $\psi(W, \theta)$, where $W \in \mathcal{W}$ denotes the relevant finite-dimensional vector of random variables in the model, and where θ is only partially identified, and constrained to lie in the identified set.

To this end, we suppose the identified set for $\theta \in \Theta$ is defined by k moment (in)equalities where the moment function $m_j(W, \theta) : \Theta \to \mathbb{R}$ is linear in $\theta \in \Theta$ for $j = 1, \ldots, k$. In this case, the identified set $\Theta_I(P)$—indexed here by the true asymptotic distribution P—is compact, and so the image of $\Theta_I(P)$ under any continuous functional $P\psi(W, \theta) : \Theta \to \mathbb{R}$ will be an interval $\Psi_I(P) = [\Psi_I^{\ell b}(P), \Psi_I^{ub}(P)]$. In this framework, the endpoints of the interval $\Psi_I(P)$ can be determined by solving two linear optimization problems:

(i) minimize $P\psi(W, \theta)$ over $\theta \in \Theta_I(P)$ to determine $\Psi_I^{\ell b}(P)$,

(ii) maximize $P\psi(W, \theta)$ over $\theta \in \Theta_I(P)$ to determine $\Psi_I^{ub}(P)$.

Seen in this way, $\Psi_I^{\ell b}(P)$ and $\Psi_I^{ub}(P)$ are the *value functions* of two stochastic linear optimization problems. Now let $\theta_0 \in \Theta$ denote the true value of the parameter, and consider the problem of constructing a confidence set $C_n^{\psi}(1-\alpha)$ that asymptotically covers $P\psi(W, \theta_0)$ with probability at least $1-\alpha$ uniformly over $(\theta, P) \in \{(\theta, P) : \theta \in \Theta_I(P), P \in \mathcal{P}\}$, where $\mathcal{P}$ is some large class of data generating processes (DGPs). In particular, we wish to construct a confidence set $C_n^{\psi}(1-\alpha)$ such that

$$\liminf_{n\to\infty} \inf_{\{(\psi,P):\ \psi\in\Psi_I(P),\ P\in\mathcal{P}\}} P(\psi \in C_n^{\psi}(1-\alpha)) \geq 1-\alpha.$$

To construct such a set, we will approximate the distribution of the endpoints $(\Psi_I^{\ell b}(P), \Psi_I^{ub}(P))$ of the identified set $\Psi_I(P)$. In particular, let $\mathcal{F}$ denote the relevant class of functions (we define this class more precisely in Section 2.3). We show that under a constraint qualification condition, for any sequence $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$ converging to a measure $P \in \mathcal{P}$ in an appropriate sense (to be made precise), there exist continuous functionals $(\Psi_I^{\ell b})'_P, (\Psi_I^{ub})'_P : \ell^{\infty}(\mathcal{F}) \to \mathbb{R}$ such that

$$\sqrt{n}\left(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)\right) \rightsquigarrow (\Psi_I^{\ell b})'_P(\mathbb{G}_P), \tag{2.1}$$

$$\sqrt{n}\left(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P_n)\right) \rightsquigarrow (\Psi_I^{ub})'_P(\mathbb{G}_P), \tag{2.2}$$

where $\mathbb{G}_P \in \ell^{\infty}(\mathcal{F})$ is the limit of the empirical process $\mathbb{G}_{n,P_n} := \sqrt{n}(\mathbb{P}_n - P_n) \in \ell^{\infty}(\mathcal{F})$, and $(\Psi_I^{\ell b}(\mathbb{P}_n), \Psi_I^{ub}(\mathbb{P}_n))$ are suitable estimates of the value functions. Moreover, we show conditions under which:

$$\sqrt{n}\left(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)\right) | \{W_i\}_{i=1}^n \rightsquigarrow (\Psi_I^{\ell b})'_P(\mathbb{G}_P), \tag{2.3}$$

$$\sqrt{n}\left(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)\right) | \{W_i\}_{i=1}^n \rightsquigarrow (\Psi_I^{ub})'_P(\mathbb{G}_P), \tag{2.4}$$

uniformly over $\mathcal{P}$, where $\mathbb{P}_n^b$ is the empirical bootstrap distribution. From here, our proposed confidence set

takes the form:

$$C_n^{\psi}(1-\alpha) := \left[\Psi_I^{\ell b}(\mathbb{P}_n) - \frac{\hat{\Psi}_{\alpha}^{\ell b}}{\sqrt{n}}, \Psi_I^{ub}(\mathbb{P}_n) + \frac{\hat{\Psi}_{\alpha}^{ub}}{\sqrt{n}}\right],$$

where the quantiles $\hat{\Psi}_{\alpha}^{\ell b}$ and $\hat{\Psi}_{\alpha}^{ub}$ are selected from the bootstrap approximation to the distributions of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$ given by (2.3) and (2.4) in order to guarantee uniformly correct coverage. Note that the approximations in (2.3) and (2.4) can be computed by repeatedly bootstrapping linear programs, motivating our "naive" bootstrap procedure. After presenting some motivating examples, the next sections develop this methodology rigorously. The general development of the methodology takes place in two parts: first, we show the conditions under which the value functions of a linear program are uniformly Hadamard directionally differentiable, and then we prove that our naive bootstrap can approximate these directional derivatives uniformly.

2.2.2 Examples

We now present some motivating examples that illustrate why inference procedures for functionals of partially identified parameters are needed.

Example 1 (Missing Data)**.** *Consider the canonical missing data example. In this example the researcher observes a sample* $\{Y_i D_i, D_i\}_{i=1}^n$*. For simplicity, suppose that* $Y_i, D_i \in \{0,1\}$*. The parameter of interest is the unconditional average of the outcome variable:*

$$P\psi(W,\theta) = \psi(\theta) = \sum_y \sum_d \theta_{yd} \cdot y,$$

where $\theta_{yd} := P(Y = y, D = d)$*. The constraints imposed by the observed distribution* $\mathbb{P}_n(YD, D)$ *on the latent distribution* $\theta_{yd} = P(Y = y, D = d)$ *are given by:*

$$\mathbb{P}_n(YD = 0, D = 1) = \theta_{01}, \tag{2.5}$$
$$\mathbb{P}_n(YD = 1, D = 1) = \theta_{11}, \tag{2.6}$$
$$\mathbb{P}_n(YD = 0, D = 0) = \theta_{00} + \theta_{10}. \tag{2.7}$$

It is straightforward to see that point identification of θ *occurs only when* $\mathbb{P}_n(D = 0) = 0$*. The identified set for our function of interest,* $\Psi_I(\mathbb{P}_n) = [\Psi_I^{\ell b}(\mathbb{P}_n), \Psi_I^{ub}(\mathbb{P}_n)]$ *can be obtained by solving the problems:*

$$\Psi_I^{\ell b}(\mathbb{P}_n) = \min_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \qquad \Psi_I^{ub}(\mathbb{P}_n) = \max_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \tag{2.8}$$

where $\Theta_I(\mathbb{P}_n)$ *is the set of* θ_{yd} *satisfying the constraints* (2.5)-(2.7)*. Note that the optimization problems in* (2.8) *are linear programs. This chapter will attempt to exploit the structure of the optimization problems in* (2.8) *to propose an inference procedure that is easy to use for functionals of partially identified parameters. Here, note that* ψ *is a functional of the partially identified parameter* θ*, where the identified set for* θ *is given by* $\Theta_I(\mathbb{P}_n)$*.*

Example 2 (Linear Regression with Interval-Valued Dependent Variable)**.** *Consider the example of linear regression with interval-valued dependent variable. We will follow closely the exposition in Kaido et al. (2019a) Appendix C. In this example the model is given by* $Y = X^T\theta + \varepsilon$*, where* $X \in \mathbb{R}^d$ *with* R *points of support. However, it is assumed that the dependent variable is interval-valued in the following way: although the value of* Y *is never observed, there exists two observable random variables* Y^* *and* Y_* *such that*

$P(Y_* \leq Y \leq Y^*) = 1$. *The objective is then to construct bounds on the parameter* θ *given that researcher observes a sample* $\{Y_i^*, Y_{*i}, X_i\}_{i=1}^n$, *and never directly observes the value of* Y. *Denoting the support points of* X *as* $\{x_1, \ldots, x_r, \ldots, x_R\}$, *as in Kaido et al. (2019a) the identified set is given by:*

$$\Theta_I(P) := \{\theta : \mathbb{E}[Y_*|X = x_r] - x_r^T\theta \leq 0,\ x_r^T\theta - \mathbb{E}[Y^*|X = x_r] \leq 0,\ r = 1, \ldots, R\}.$$

We now suppose that the researcher is interested in conducting inference only on the first component θ_1 *of the parameter vector* θ. *Then in our notation we can set* $\psi(W, \theta) = \psi(\theta) = \theta_1$. *Under some weak conditions we will have that the identified set for the functional* ψ *is an interval* $\Psi_I(\mathbb{P}_n) = [\Psi_I^{\ell b}(\mathbb{P}_n), \Psi_I^{ub}(\mathbb{P}_n)]$ *with the endpoints determined by the program:*

$$\Psi_I^{\ell b}(\mathbb{P}_n) = \min_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \qquad \Psi_I^{ub}(\mathbb{P}_n) = \max_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \tag{2.9}$$

where $\Theta_I(\mathbb{P}_n)$ *is the estimate of the identified set obtained by replacing the moment conditions with their sample analogs. Note that since all moment conditions defining the identified set are linear in* θ, *the optimization problems in* (2.9) *are linear programs. Again, this chapter will propose an inference procedure for functionals of partially identified parameters that uses the special structure of the optimization problems in* (2.9) *that characterizes the functional bounding problem.*

Example 3 (Nonparametric State Dependence)**.** *Consider the model of nonparametric state dependence given in Torgovitsky (2016). In this model, the researcher observes a realization of a random sequence* $Y := (Y_0, \ldots, Y_T)$ *for each individual for* T *periods. As in Torgovitsky (2016), we consider for simplicity that each outcome* Y_t *is binary, so that* $Y \in \{0, 1\}^{T+1}$. *The sequence of observed outcomes* Y *are related to a sequence of unobserved potential outcomes* $U(0) := (U_1(0), \ldots, U_T(0))$ *and* $U(1) := (U_1(1), \ldots, U_T(1))$ *through the equation:*

$$Y_t = Y_{t-1}U_t(0) + (1 - Y_{t-1})U_t(1).$$

The researcher may also have access to a sequence of covariates $X := (X_0, \ldots, X_T)$ *for each individual. The object of interest for the researcher is assumed to be treatment effect parameters that depend on the unobserved potential outcomes* $(U_t(0), U_t(1))$ *at time* $1 \leq t \leq T$. *Examples of such treatment effect parameters include the average treatment effect, given by* $ATE_t = P(U_t(0) = 0, U_t(1) = 1) - P(U_t(0) = 1, U_t(1) = 0)$, *or the voting criterion given by* $P(U_t(0) = 0, U_t(1) = 1)$ *(or* $P(U_t(0) = 1, U_t(1) = 0)$*).*

To see how to bound these parameters, define the vector

$$\mathbf{u} := (u_0, u_1(0), \ldots, u_T(0), u_1(1), \ldots, u_T(1))',$$

where u_0 *is the initial (period 0) potential outcome. In addition, let* $U := (U_0, U(0), U(1))'$, *and let*

$$\mathcal{U}^\dagger(\mathbf{y}) := \{\mathbf{u} :\ u_0 = y_0,\ y_t = y_{(t-1)}u_t(0) + (1 - y_{(t-1)})u_t(1),\ \forall t\},$$

which is the set of all vectors $\mathbf{u}$ *of potential outcomes that could rationalize an observed vector of outcomes* $\mathbf{y} = (y_0, \ldots, y_T)'$. *Finally, let* $\mathbf{X} = (x_0, \ldots, x_T)'$. *Torgovitsky (2016) shows that without any additional restrictions, the sharp set of constraints on the unobserved joint distribution* $\theta_{\mathbf{u},\mathbf{x}} := P(U = \mathbf{u}, X = \mathbf{x})$ *is given by:*

$$\mathbb{P}_n(Y = \mathbf{y}, X = \mathbf{x}) = \sum_{\mathbf{u} \in \mathcal{U}^\dagger(\mathbf{y})} \theta_{\mathbf{u},\mathbf{x}}. \tag{2.10}$$

Torgovitsky (2016) shows how additional restrictions can also be imposed on the unobserved joint distribution $\theta_{\mathbf{u},\mathbf{x}}$, such as monotone treatment response (MTR) constraints, stationarity (ST) constraints, monotone instrumental variable (MIV) constraints and monotone treatment selection (MTS) constraints. All of these constraints can be imposed on the optimization problem as moment-inequality constraints. Let $\Theta_I(\mathbb{P}_n)$ denote the set of all joint distributions θ satisfying the imposed constraints as well as the observational equivalence condition (2.10). Proposition 1 in Torgovitsky (2016) shows that if $\psi : \Theta_I(\mathbb{P}_n) \to \mathbb{R}$ is a continuous treatment effect parameter, then the identified set for ψ can be estimated by $\Psi_I(\mathbb{P}_n) = [\Psi_I^{\ell b}(\mathbb{P}_n), \Psi_I^{ub}(\mathbb{P}_n)]$, and can be obtained by solving the problems:

$$\Psi_I^{\ell b}(\mathbb{P}_n) = \min_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \qquad \Psi_I^{ub}(\mathbb{P}_n) = \max_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta). \tag{2.11}$$

Note that when T is large, there can be a large number of constraints defining the set $\Theta_I(\mathbb{P}_n)$, and the partially identified parameter θ can be high-dimensional.

Example 4 (Inference on Counterfactual Policies)**.** *In the setting of Kasy (2016), the researcher is interested in ranking counterfactual policies "A" and "B" which represent two competing proposals of assigning individuals to some treatment based on covariate values. It is assumed that the policy maker only has knowledge of the partially-identified parameters $g_0(X) := \mathbb{E}[Y_0|X]$ and $g_1(X) := \mathbb{E}[Y_1|X]$, where Y_d is the partially-observed potential outcome for treatment state $D = d$.*

We assume that the researcher's object of interest is the linear functional $\psi := \psi(f^A, f^B)$ where f^A is the distribution of the random variable Y^A representing the observed outcome under policy A, and f^B is the distribution of the random variable Y^B representing the observed outcome under policy B. Furthermore, let D^A be the random variable representing treatment assignment under policy A, and let D^B be the random variable representing treatment under assignment B, and assume that $D^A, D^B \perp\!\!\!\perp (Y_0, Y_1)|X$. Some simple objective functions include $\psi^A := \mathbb{E}[Y^A]$ (or $\psi^B := \mathbb{E}[Y^B]$), which measures the average outcome under policy A, or $\psi^{AB} := \mathbb{E}[Y^A - Y^B]$, which measures the difference in average outcomes between policies A and B. Let $\mathcal{G}_d$ denote the identified set for $g_d(X)$. Note that the objective function ψ^A can be decomposed as:

$$\begin{aligned}
\psi^A &= \mathbb{E}[Y^A] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y^A|D^A = 1, X\right] P(D^A = 1|X) + \mathbb{E}\left[Y^A|D^A = 0, X\right](1 - P(D^A = 1|X))\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y_0|X] + P(D^A = 1|X)\left(\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]\right)\right] \\
&= \mathbb{E}\left[g_0(X) + h^A(X)\left(g_1(X) - g_0(X)\right)\right],
\end{aligned}$$

where $h^A(X) = P(D^A = 1|X)$. Since $g_0(\cdot)$ and $g_1(\cdot)$ are only partially-identified, ψ^A will also be partially identified. Let $\Psi_I^A(P) = [\Psi_{\ell b}^A(P), \Psi_{ub}^A(P)]$ denote the identified set for $\psi^A = E[Y^A]$, where the endpoints of Ψ_I^A are determined by:

$$\Psi_{\ell b}^A(P) = \inf_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} \left[g_0(x) + h^A(x)\left(g_1(x) - g_0(x)\right)\right] P(X = x), \tag{2.12}$$

$$\Psi_{ub}^A(P) = \sup_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} \left[g_0(x) + h^A(x)\left(g_1(x) - g_0(x)\right)\right] P(X = x), \tag{2.13}$$

where $P(X = x)$ is the probability $X = x$ in the target population. Similarly, as in Kasy (2016), the objective function ψ^{AB} can be decomposed as:

$$\begin{aligned}
\psi^{AB} &= \mathbb{E}\left[Y^A - Y^B\right] \\
&= \mathbb{E}\left[\left(h^A(X) - h^B(X)\right)(Y_1 - Y_0)\right]
\end{aligned}$$

$$= \mathbb{E}\left[h^{AB}(X)g(X)\right],$$

where $h^{AB}(X) = h^A(X) - h^B(X)$, $h^A(X) = P(D^A = 1|X)$, $h^B(X) = P(D^B = 1|X)$ *and* $g(X) = g_1(X) - g_0(X)$. *Since* $g(X)$ *is only partially identified, the objective function* ψ^{AB} *will also only be partially identified. Let* $\Psi_I^{AB}(P) = [\Psi_{\ell b}^{AB}(P), \Psi_{ub}^{AB}(P)]$ *denote the identified set for* ψ^{AB}, *where the endpoints of* Ψ_I^{AB} *are given by:*

$$\Psi_{\ell b}^{AB}(P) = \inf_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} h^{AB}(x)\left(g_1(x) - g_0(x)\right)P(X = x), \tag{2.14}$$

$$\Psi_{ub}^{AB}(P) = \sup_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} h^{AB}(x)\left(g_1(x) - g_0(x)\right)P(X = x), \tag{2.15}$$

where $P(X = x)$ *is the probability* $X = x$ *in the target population. In this example, note that the partially identified parameter is* $\theta = (g_0, g_1)$ *and the identified set is* $\Theta_I(P) = \mathcal{G}_0 \times \mathcal{G}_1$.

Remark 2.2.1. *In practice, the probabilities* $P(X = x)$ *in the optimization problems* (2.12) *and* (2.13), *or* (2.14) *and* (2.15), *may need to be estimated, meaning that the objective functions in these optimization problems contain sampling uncertainty that must be accounted for when performing inference on either* Ψ_I^A *or* Ψ_I^{AB} *in addition to the sampling uncertainty inherent in the estimation of the sets* $\mathcal{G}_0$ *and* $\mathcal{G}_1$. *Currently, we are unaware of any uniformly valid inference procedure in partially identified models that can handle these cases.*

2.3 Methodology

In this section, we develop the ideas introduced in the previous section. We consider a setting where the identified set $\Theta_I(P)$ is defined by moment equalities and inequalities that are satisfied at the true parameter θ_0:

$$Pm_j(W, \theta_0) = 0, \qquad \text{for } j = 1, \ldots, r_1, \tag{2.16}$$

$$Pm_j(W, \theta_0) \leq 0, \qquad \text{for } j = r_1 + 1, \ldots, r_1 + r_2. \tag{2.17}$$

Note that we can always convert these moment equalities/inequalities defined above into $k = 2r_1 + r_2$ equivalent moment inequalities given by:

$$Pm_j(W, \theta_0) \leq 0, \qquad \text{for } j = 1, \ldots, r_1, \tag{2.18}$$

$$-Pm_j(W, \theta_0) \leq 0, \qquad \text{for } j = 1, \ldots, r_1, \tag{2.19}$$

$$Pm_j(W, \theta_0) \leq 0, \qquad \text{for } j = r_1 + 1, \ldots, r_1 + r_2. \tag{2.20}$$

Thus, we will assume throughout most of the exposition that the model is defined only by k moment inequalities:

$$Pm_j(W, \theta_0) \leq 0, \qquad \text{for } j = 1, \ldots, k. \tag{2.21}$$

Only on rare occasions will it be necessary to know which of the moment inequalities correspond to moment equalities; in these cases we will simply refer back to the original formulation in (2.16) and (2.17).

We assume that the researcher is interested in bounding the expected value of a function $\psi : \mathcal{W} \times \Theta \to \mathbb{R}$.

Define the following class of functions:

$$\mathcal{F} := \left\{ (\psi(W,\theta), m_1(W,\theta), \ldots, m_k(W,\theta))^T : \theta \in \Theta \right\}. \tag{2.22}$$

A typical element of $\mathcal{F}$ will then be the vector-valued function:

$$f(W,\theta) = \left[\psi(W,\theta), m_1(W,\theta), \ldots, m_k(W,\theta) \right]^T.$$

Furthermore, we will equip this class of functions with a semimetric that depends on the probability measure P:

$$\rho_P(\theta,\theta') := \left|\left| \text{diag}(V_P \left(f(W,\theta) - f(W,\theta') \right)^{1/2}) \right|\right|, \tag{2.23}$$

for $\theta, \theta' \in \Theta$.[2] This semimetric was also considered in Bugni et al. (2015). Furthermore, define the class:

$$\mathcal{F}' := \{ f - f' : f, f' \in \mathcal{F} \},$$

and let $\mathcal{G} = \mathcal{F} \cup \mathcal{F}^2 \cup (\mathcal{F}')^2$. Let $\{P_n \in \mathcal{P}\}_{n \geq 1}$ be any sequence of data-generating measures. Throughout the text, we will interpret the statement $P_n \rightsquigarrow P$ as weak convergence of the measures $\{P_n \in \mathcal{P}\}_{n \geq 1}$ to $P \in \mathcal{P}$; that is, weak* convergence. Asymptotically validity of our confidence set uniformly over $\mathcal{P}$ is equivalent to asymptotic validity along any sequence $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$. Under the assumptions to be presented shortly, the collection $\mathcal{P}$ will be closed and uniformly tight. Thus, from any sequence $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$ we can extract a weakly convergent subsequence converging to some $P \in \mathcal{P}$. As a result, for our purposes we will have that asymptotically validity of our proposed confidence set uniformly over $\mathcal{P}$ will be equivalent to asymptotic validity along any weakly converging sequence $P_n \rightsquigarrow P$. The reader is encouraged to keep this in mind throughout.

Now, let $\mathscr{P}$ denote the collection of all probability measures on $\mathcal{W}$. We now impose the following assumptions:

Assumption 2.3.1. *The parameter space $(\Theta, \mathcal{P})$ satisfies the following conditions:*

(i) $\Theta \subset \mathbb{R}^{d_\theta}$ *is convex and compact.*

(ii) $\mathcal{F}$ *is a measurable class of functions.*

(iii) Each distribution $P \in \mathcal{P} \subseteq \mathscr{P}$ *satisfies:*

- *(a)* $Pm_j(W,\theta_0) \leq 0$, *for* $j = 1, \ldots, k$.
- *(b) In a sample* $\{W_i\}_{i=1}^n$, W_i *are independent and identically distributed according to* $P \in \mathcal{P}$.

(iv) There exists a bounded envelope function F *for the class* $\mathcal{F}$ *such that for some* $a > 0$,

$$\sup_{P \in \mathcal{P}} \max \left\{ P||F(W)||^{2+a}, PW \right\} < \infty.$$

Remark 2.3.1. *In Assumption 2.3.1(ii), we call* $\mathcal{F}$ *measurable if* $\mathcal{F}$ *is* $P-$*measurable in the sense of Van Der Vaart and Wellner (1996) Definition 2.3.3 for all probability measures* $P \in \mathcal{P}$.

[2] Recall a semimetric satisfies (i) $\rho(f,f) = 0$, (ii) $\rho(f,g) \leq \rho(f,h) + \rho(h,g)$ and (iii) $\rho(x,y) = \rho(y,x)$. However, unlike a metric, a semimetric can be equal to zero when evaluated at two distinct elements.

Note that we can write the identified set $\Theta_I(P)$ as:

$$\Theta_I(P) = \{\theta \in \Theta : Pm_j(W, \theta) \leq 0, \quad j = 1, \ldots, k\}. \tag{2.24}$$

Now let $\Theta_I(\mathbb{P}_n)$ denote the estimate of the identified set:

$$\Theta_I(\mathbb{P}_n) = \{\theta \in \Theta : \mathbb{P}_n m_j(W, \theta) \leq 0, \quad j = 1, \ldots, k\}, \tag{2.25}$$

where $\mathbb{P}_n$ denotes the empirical measure for the first n observations:

$$\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{W_i}, \tag{2.26}$$

where δ_{W_i} is the Dirac delta function. We restrict attention to a certain class of functionals of the identified set.

Assumption 2.3.2. *(i) The functional of interest $\psi(w, \theta) : \mathcal{W} \times \Theta \to \mathbb{R}$ is linear in θ, and is continuous in $w \in \mathcal{W}$; (ii) the functions $m_j(w, \theta) : \mathcal{W} \times \Theta \to \mathbb{R}$ are linear in θ and continuous in $w \in \mathcal{W}$ for $j = 1, \ldots, k$.*

Denote the identified set for $P\psi(W, \theta)$ as $\Psi_I(P)$, and note that the identified set for $P\psi(W, \theta)$ is the projection of $\Theta_I(P)$ on the manifold generated by $P\psi(W, \theta)$. As such, under standard conditions (see Lemma 2.A.4(ii)) the projection estimator $\Psi_I(\mathbb{P}_n)$ will be a consistent estimator of $\Psi_I(P)$. Moreover, since $P\psi(W, \theta)$ is continuous and $\Theta_I(P)$ is convex and compact, the identified set $\Psi_I(P)$ is an interval—$\Psi_I(P) = [\Psi_I^{\ell b}(P), \Psi_I^{ub}(P)]$—with endpoints determined by:

$$\Psi_I^{\ell b}(P) := \inf_{\theta \in \Theta} \quad P\psi(W, \theta) \qquad \text{s.t.} \qquad Pm_j(W, \theta) \leq 0, \qquad j = 1, \ldots, k, \tag{2.27}$$

$$\Psi_I^{ub}(P) := \sup_{\theta \in \Theta} \quad P\psi(W, \theta) \qquad \text{s.t.} \qquad Pm_j(W, \theta) \leq 0, \qquad j = 1, \ldots, k. \tag{2.28}$$

However, since P is not known, the programs (2.27) and (2.28) will be approximated using the empirical distribution $\mathbb{P}_n$ by replacing the population moment conditions and objective function with their sample counterparts:

$$\Psi_I^{\ell b}(\mathbb{P}_n) := \inf_{\theta \in \Theta} \quad \mathbb{P}_n\psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq 0, \qquad j = 1, \ldots, k, \tag{2.29}$$

$$\Psi_I^{ub}(\mathbb{P}_n) := \sup_{\theta \in \Theta} \quad \mathbb{P}_n\psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq 0, \qquad j = 1, \ldots, k. \tag{2.30}$$

After an estimate of the identified set is obtained, interest will lie in constructing uniformly valid confidence sets for the true parameter $\psi_0 := P\psi(W, \theta_0)$. To perform inference on the true parameter using the optimal values $(\Psi_I^{\ell b}(P), \Psi_I^{ub}(P))$ in programs (2.27) and (2.28), we will approximate the distributions of $\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P))$ and $\sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P))$ by a simple bootstrap procedure, and will be particularly interested in proving the procedure is valid *uniformly over* $\mathcal{P}$.

Remark 2.3.2. *As a technical note, the functions $(\Psi_I^{\ell b}(\cdot), \Psi_I^{ub}(\cdot))$ will be seen as maps from $\mathcal{P}_+$ to $\mathbb{R}$, where $\mathcal{P}_+$ is defined as the collection of all measures $\mathcal{P}$ as well as all finite empirical measures $\mathbb{P}_n$ generated by a $P \in \mathcal{P}$ (i.e. $\mathcal{P}_+ = span(\mathcal{P}, \{\delta_w\}_{w \in \mathcal{W}})$, where $\{\delta_w\}_{w \in \mathcal{W}}$ is any finite collection of point masses). It will be useful to distinguish between the collections $\mathcal{P}$ and $\mathcal{P}_+$ throughout.*

2.3.1 Value Function Differentiability

Recall from the discussion in Section 2.2 that our first step will be to show that the value functions satisfy an appropriate level of differentiability with respect to the underlying probability measure. Since the underlying probability measure is a possibly infinite-dimensional object, we must use a form of differentiability that is valid between metric spaces. In particular, it is well-known (e.g. Shapiro (1990), Shapiro (1991)) that under some conditions the functions $(\Psi_I^{\ell b}(P), \Psi_I^{ub}(P))$ are *Hadamard directionally differentiable*. To introduce the differentiability concepts used in this chapter in general form, let $\mathbb{D}$ and $\mathbb{E}$ be topological vector spaces.

Definition 2.3.1 (Hadamard Directional Differentiability)**.** *A map $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is called Hadamard directionally differentiable at $\zeta \in \mathbb{D}_\phi$ if there is a linear map $\phi'_\zeta : \mathbb{D} \to \mathbb{E}$ such that*

$$\frac{\phi(\zeta + t_n h_n) - \phi(\zeta)}{t_n} \to \phi'_\zeta(h),$$

for converging sequences $\{t_n\} \subset \mathbb{R}_+$ with $t_n \downarrow 0$ and $h_n \to h$ such that $\zeta + t_n h_n \in \mathbb{D}_\phi$ for every n. In addition, we say ϕ is Hadamard directionally differentiable tangential to a set $\mathbb{D}_0 \subseteq \mathbb{D}$ if we also require that the limit $h \in \mathbb{D}_0$ in the above.

While Hadamard directional differentiability can be used to justify an inference procedure in stochastic programs for a fixed data-generating measure $P \in \mathcal{P}$ (c.f. Shapiro (1991)), it is not sufficient to construct an inference procedure for stochastic programs that is valid *uniformly* over $\mathcal{P}$. It is natural to wonder whether stochastic programs are uniformly Hadamard directionally differentiable, which is defined in the following:

Definition 2.3.2 (Uniform Hadamard Directional Differentiability)**.** *Let $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$, $\mathbb{D}_0 \subseteq \mathbb{D}$, and $\mathbb{D}_\zeta \subseteq \mathbb{D}_\phi$. The map $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is called uniformly Hadamard directionally differentiable in $\zeta \in \mathbb{D}_\zeta$ if there is a continuous map $\phi'_\zeta : \mathbb{D} \to \mathbb{E}$ such that*

$$\frac{\phi(\zeta_n + t_n h_n) - \phi(\zeta_n)}{t_n} \to \phi'_\zeta(h), \tag{2.31}$$

for all converging sequences $\zeta_n \to \zeta \in \mathbb{D}_\zeta$, $\{t_n\} \subset \mathbb{R}_+$ with $t_n \downarrow 0$, and $h_n \to h$ such that $\zeta_n + t_n h_n \in \mathbb{D}_\phi$ for every n. In addition, we say ϕ is uniformly Hadamard directionally differentiable tangential to a set $\mathbb{D}_0 \subseteq \mathbb{D}$ if we also require that the limit $h \in \mathbb{D}_0$ in the above.

This definition is analogous to the extension of Hadamard differentiability to uniform Hadamard differentiability presented in Belloni et al. (2017), although our definition restricts $t_n \to 0$ from above (providing a "direction"). It also allows the spaces involved to be topological vector spaces rather than normed linear spaces, and allows the derivative map ϕ'_ζ to be continuous rather than linear.

In addition, reflecting more the definition in Van Der Vaart and Wellner (1996) p. 379, we do not explicitly require that the derivative map $(\zeta, h) \mapsto \phi'_\zeta(h)$ be continuous at every (ζ, h), as is done in the extension of Hadamard differentiability to uniform Hadamard differentiability in Belloni et al. (2017). However, similar to Belloni et al. (2017), we will use the flexibility provided by the above definition to allow ζ_n to lie outside $\mathbb{D}_\zeta$.

As we will see, under some conditions the value functions of a stochastic program are differentiable in the sense of Definition 2.3.2. Our first main result requires non-emptiness of the identified set, the existence and uniqueness of Lagrange multipliers, and uniqueness of the optimal solutions in the programs (2.27) and (2.28). To guarantee these properties will require that a "uniform constraint qualification" holds for the linear programs. We impose such a constraint qualification in the next assumption.

Let $\mathscr{A}(\theta, P)$ be an index set defined as:

$$\mathscr{A}(\theta, P) := \{j \in \{r_1 + 1, \ldots, r_1 + r_2\} : Pm_j(W, \theta) = 0\}, \tag{2.32}$$

i.e. $\mathscr{A}(\theta, P)$ denotes the set indexing the binding moment inequalities at θ for some probability measure $P \in \mathcal{P}$. Finally, let $d_{BL} : \mathscr{P} \times \mathscr{P} \to \mathbb{R}_+$ denote the bounded Lipschitz metric, which is defined for any metric space $\mathbb{D}$ by:

$$d_{BL}(P, Q) := \sup_{f \in BL_1(\mathbb{D})} \left| \int f \, dP - \int f \, dQ \right|,$$

where:

$$BL_1(\mathbb{D}) := \{f : \mathbb{D} \to \mathbb{R} : ||f||_\infty \leq 1 \text{ and } |f(x) - f(y)| \leq |x - y| \text{ for all } x \neq y\}.$$

Assumption 2.3.3. *Let $\theta^*_{\ell b}(P)$ and $\theta^*_{ub}(P)$ be the optimal solutions to the problems (2.27) and (2.28), let $\boldsymbol{G}(\theta, P)$ be the matrix formed by vertically stacking the row vectors $\{\nabla_\theta Pm_j(W, \theta)\}_{j=1}^{r_1}$ and $\{\nabla_\theta Pm_j(W, \theta)\}_{j \in \mathscr{A}(\theta, P)}$, and let $\mathcal{P}^\epsilon := \{Q \in \mathcal{P}_+ : d_{BL}(Q, P) \leq \varepsilon, P \in \mathcal{P}\}$. Then there exists $\varepsilon > 0$ such that:*

(i) $\Theta_I(P) \neq \emptyset$ *for all* $P \in \mathcal{P}^\epsilon$.

(ii) (LICQ) There exists a $\kappa > 0$ *such that:*

$$\inf_{P \in \mathcal{P}^\epsilon} \min \left\{ eig\left(\boldsymbol{G}(\theta^*_{\ell b}(P), P)\boldsymbol{G}(\theta^*_{\ell b}(P), P)^T\right), eig\left(\boldsymbol{G}(\theta^*_{ub}(P), P)\boldsymbol{G}(\theta^*_{ub}(P), P)^T\right) \right\} \geq \kappa. \tag{2.33}$$

where $eig(A)$ denotes the minimum eigenvalue of A.

(iii) The optimal solutions $\theta^*_{\ell b}(P)$ and $\theta^*_{ub}(P)$ are unique uniformly over $\mathcal{P}^\varepsilon$.

Assumption 2.3.3(i) implies that, for large enough n, the identified set is non-empty. Given non-emptiness of the identified set, Assumption 2.3.3(ii) implies a uniform version of the *linear independence constraint qualification* (LICQ), and is instrumental in ensuring the existence and uniqueness of Lagrange multipliers in the programs (2.27) and (2.28). Finally, the interpretation of Assumption 2.3.3(iii) is straightforward.

Constraint qualifications in various forms have appeared throughout the recent history of partial identification (e.g. Beresteanu and Molinari (2008), Pakes et al. (2011) Kaido and Santos (2014), Freyberger and Horowitz (2015), Kaido et al. (2019a), Gafarov et al. (2018), and Gafarov (2019)). We refer to the recent paper of Kaido et al. (2019b) for a full comparison of the constraint qualifications used in partial identification. There are some cases where it may be easy to directly verify that the Assumption 2.3.3(ii) is satisfied, but in general Assumption 2.3.3 is a high-level condition.[3] We shall attempt to provide some more perspective on the strength of this assumption in our discussion in Section 2.4.

All components of Assumption 2.3.3 are regularity assumptions that are important in the proof of uniform Hadamard directional differentiability. Specifically, it is necessary to restrict the optimal solutions and Lagrange multipliers in (2.27) and (2.28) to be unique for all $P \in \mathcal{P}$. To understand why, consider the problem of multiple optimal solutions, and note that if the problems (2.27) and (2.28) admit multiple solutions there may be differences between the sets representing "the limiting optimal solutions" (over the sequence $\{P_n \in \mathcal{P}\}_{n=1}^\infty$), and the sets representing "the optimal solutions at the limit" ($P \in \mathcal{P}$). This is related to the Theorem of the Maximum, and the fact that the Theorem of the Maximum guarantees only that the solution

[3] For example, if the moment functions can be expressed as $Pm_j(W, \theta) = P\tilde{m}_j(W) + a_j'\theta$, where $\tilde{m}_j$ is a function of the random variable $W \in \mathcal{W}$, and $a_j \in \mathbb{R}^d$ is a vector, then it suffices to verify the Jacobian of the moment functions (w.r.t. θ) has full column rank.

correspondence is upper hemicontinuous, but not lower hemicontinuous (and thus, not continuous). In this case it is possible to show that the value functions $\Psi_I^{\ell b}(\cdot)$ and $\Psi_I^{ub}(\cdot)$ are Hadamard directionally differentiable, but not necessarily *uniformly* Hadamard directionally differentiable. The same intuition follows for the Lagrange multipliers. However, Wachsmuth (2013) shows that the LICQ—implied by Assumption 2.3.3—is the weakest constraint qualification under which the Lagrange multipliers are guaranteed to be unique. Since the LICQ is implied uniformly over $\mathcal{P}^\varepsilon$ by Assumption 2.3.3(ii), existence and uniqueness of Lagrange multipliers also follows from Assumption 2.3.3(ii). Assumption 2.3.3(iii) then imposes uniqueness of optimal solutions separately. Assumption 2.3.3(i) enables us to impose Assumptions 2.3.3(ii) and (iii) over $\mathcal{P}^\varepsilon$, but is also required to ensure that the uniform Hadamard directional derivative is well-defined, which would not be the case if the identified set was allowed to be empty for all n.

A few additional remarks about Assumption 2.3.3 are in order. First, Assumption 2.3.3 is one of the rare cases where it is useful to distinguish between moment equalities as in (2.16) and moment inequalities as in (2.17), since the assumption imposes different conditions on the two types of moments. Next note that it is possible to show that Assumption 2.3.3 implies that at every $P \in \mathcal{P}$ there must be at least one interior point of the set defined by the moment inequalities at which all moment equalities are satisfied. The major restriction imposed by this implication is that the moment inequalities evaluated at the limiting $P \in \mathcal{P}$ cannot point-identify the parameter of interest.[4] This condition is reminiscent of condition 4 in Theorem 2.1 in Shi and Shum (2015), and its discussion on page 499 of Shi and Shum (2015). Similar to their discussion, we note that in many cases this assumption will fail when two inequality constraints become equivalent, in which case the inequality constraints can be combined to form an equality constraint so that the assumption still holds. Finally, note that this assumption is sufficient for our method to be uniformly valid, but is not necessary. However, the assumption is the most primitive assumption we are currently aware of, as it connects to the highly used constraint qualification assumptions in optimization literature while imposing minimal constraints on any sequence $\{P_n \in \mathcal{P}_+\}_{n=1}^\infty$ required for uniformity.

The final assumption relates to the gradient of the objective function and the moments:

Assumption 2.3.4. *The gradients* $\{\nabla_\theta P\psi(W,\theta), \{\nabla_\theta Pm_j(W,\theta)\}_{j=1}^k\}$ *are uniformly bounded over* $\mathcal{P}^\varepsilon$.

This assumption is required only to show that the Lagrange multipliers are uniformly bounded over $\mathcal{P}^\varepsilon$. Any other assumption that implies uniform boundedness of the Lagrange multipliers might then be safely substituted for Assumption 2.3.4.

Finally, as a piece of technical machinery, we define the tangent cone as:

$$\mathcal{T}_P(\mathcal{F}) = \{v \in UC_b(\mathcal{F}, \rho_P) : \forall t_n \downarrow 0,\ \forall\{P_n \in \mathcal{P}\}_{n=1}^\infty \rightsquigarrow P \in \mathcal{P},\ \exists\{Q_n \in \mathcal{P}_+\}_{n=1}^\infty \text{ s.t. } t_n^{-1}(Q_n - P_n) \to v\}, \tag{2.34}$$

where $UC_b(\mathcal{F}, \rho_P) \subset \ell^\infty(\mathcal{F})$ denotes the space of bounded, and uniformly continuous functions with respect to the semimetric ρ_P defined in (2.23). While restricting the tangent cone to be a subset of $UC_b(\mathcal{F}, \rho_P)$ might appear to be restrictive, under the Donsker-type assumptions to be introduced later almost all paths of the limit of the empirical process $\sqrt{n}(\mathbb{P}_n - P_n)$ will be uniformly continuous; see Addendum 1.5.8 in Van Der Vaart and Wellner (1996). We now have the following result:

Theorem 2.3.1. *Suppose Assumptions 2.3.1 - 2.3.4 hold, and consider* $\Psi_I^{\ell b}, \Psi_I^{ub} : \mathcal{P}_+ \to \mathbb{R}$ *defined by the programs (2.29) and (2.30). Then* $\Psi_I^{\ell b}(\cdot), \Psi_I^{ub}(\cdot)$ *are uniformly Hadamard directionally differentiable tangential to* $\mathcal{T}_P(\mathcal{F})$. *In particular, for all weakly converging sequences* $P_n \rightsquigarrow P \in \mathcal{P}$, $\{t_n\} \subset \mathbb{R}_+$ *with* $t_n \downarrow 0$,

[4] Note also that this condition rules out the case that the moment inequalities define an empty region. However, we do not consider this a "major restriction" of our method, since if the true identified set is empty then computing functionals over the identified set becomes a dubious exercise.

and $h_n \to h \in \mathcal{T}_P(\mathcal{F})$ *such that* $P_n + t_n h_n \in \mathcal{P}_+$ *for every* n*, we have:*

$$(\Psi_I^{\ell b})'_P(h) := \lim_{n\to\infty} \frac{\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n)}{t_n} = h_1 \psi(W, \theta^*_{\ell b}(P)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P) h_{j+1} m_j(W, \theta^*_{\ell b}(P))), \quad (2.35)$$

$$(\Psi_I^{ub})'_P(h) := \lim_{n\to\infty} \frac{\Psi_I^{ub}(P_n + t_n h_n) - \Psi_I^{ub}(P_n)}{t_n} = -h_1 \psi(W, \theta^*_{ub}(P)) + \sum_{j=1}^{k} \lambda^*_{ub,j}(P) h_{j+1} m_j(W, \theta^*_{ub}(P))), \quad (2.36)$$

where $h_j f_j$ *is the* j^{th} *component of* hf *for* $f \in \mathcal{F}$*,* $\theta^*_{\ell b}(P)$ *and* $\theta^*_{ub}(P)$ *are the optimal solutions in the lower and upper bounding problems at* $P \in \mathcal{P}$*, and* $\{\lambda^*_{\ell b,j}(P)\}_{j=1}^k$ *and* $\{\lambda^*_{ub,j}(P)\}_{j=1}^k$ *are the Lagrange multipliers in the lower and upper bounding problems at* $P \in \mathcal{P}$*.*

The uniform component of this theorem lies in the fact that it is valid over any generating sequence $\{P_n \in \mathcal{P}\} \rightsquigarrow P \in \mathcal{P}$. This uniform version of differentiability turns out to be sufficient to apply the extended continuous mapping theorem (Theorem 1.11.1 in Van Der Vaart and Wellner (1996)) in order to relate this result to inference on the optimal value functions. This is exactly what is done in Lemma 2.3.1 in the next subsection.

2.3.2 From Differentiability to Weak Convergence

We now consider the asymptotic distribution of the properly rescaled and recentered value functions given in (2.29) and (2.30), which will make use of the uniform differentiability property given in Theorem 2.3.1. To cover the case of a drifting sequence of data-generating processes, which will be necessary to show uniformity, we impose additional assumptions.

Assumption 2.3.5. *The collections* $\mathcal{F}$ *and* $\mathcal{P}$ *satisfy the following:*

(i) The empirical process $\mathbb{G}_{n,P} := \sqrt{n}(\mathbb{P}_n - P)$ *is asymptotically equicontinuous uniformly over* $\mathcal{P}$*; that is, for every* $\varepsilon > 0$*,*

$$\lim_{\delta \downarrow 0} \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P^*_P \left(\sup_{\rho_P(\theta,\theta')<\delta} ||\mathbb{G}_{n,P} f(W,\theta) - \mathbb{G}_{n,P} f(W,\theta')|| > \varepsilon \right) = 0,$$

where ρ_P *is as in* (2.23)*.*

(ii) The semimetric ρ_P *satisfies:*

$$\lim_{\delta \downarrow 0} \sup_{||(\theta_1,\theta_1')-(\theta_2,\theta_2')||<\delta} \sup_{P\in\mathcal{P}} |\rho_P(\theta_1,\theta_1') - \rho_P(\theta_2,\theta_2')| = 0.$$

(iii) Let $\mathscr{A}(\theta, P) \subseteq \{r_1 + 1, \ldots, r_1 + r_2\}$ *denote the binding moment inequalities at* (θ, P)*, let* $\mathcal{I}_{eq} = \{1, \ldots, r_1\}$*, and let* $V_j(\theta) := Var(m_j(W, \theta))$*, for* $j = 1, \ldots, r_1 + r_2$*. Then there exists a constant* $\underline{v} > 0$ *such that for all* $P \in \mathcal{P}$*:*

$$\inf_{\theta\in\Theta_I(P)} \min_{j\in\mathscr{A}(\theta,P)\cup\mathcal{I}_{eq}} V_j(\theta). \geq \underline{v}$$

(iv) Let $\mathscr{A}(\theta, P) \subseteq \{r_1 + 1, \ldots, r_1 + r_2\}$ *denote the binding moment inequalities at* (θ, P)*, and let* $\underline{\sigma} > 0$ *be a constant. One of the following two holds:*

(a) Let $\boldsymbol{V}_P^m := Var_P\left[\{m_j(W,\theta)\}_{j=1}^{r_1}, \{m_j(W,\theta)\}_{j\in\mathscr{A}(\theta,P)}\right]$. *The objective function* $\psi(w,\theta)$ *is a trivial function of* w, *and we have:*

$$\inf_{\theta\in\Theta_I(P)} eig(\boldsymbol{V}_P^m) \geq \underline{\sigma}.$$

(b) Let $\boldsymbol{V}_P := Var_P\left[\psi(W,\theta), \{m_j(W,\theta)\}_{j=1}^{r_1}, \{m_j(W,\theta)\}_{j\in\mathscr{A}(\theta,P)}\right]$. *Then we have:*

$$\inf_{\theta\in\Theta_I(P)} eig(\boldsymbol{V}_P) \geq \underline{\sigma}.$$

(v) There exist positive constants $C, \delta > 0$ *such that* $\max_{j=1,\ldots,k} |Pm_j(W,\theta)| \geq C\min(\delta, d_H(\theta,\Theta_I(P)))$ *for every* $P\in\mathcal{P}$ *and* $\theta\in\Theta$, *where* d_H *is the Hausdorff metric.*

Assumptions 2.3.5(i) and 2.3.5(ii) and are required to apply a uniform Donsker theorem to the class of functions $\mathcal{F}$. Also related are Assumption 2.3.5(iii) and 2.3.5(iv), which are required to ensure a uniform multivariate central limit theorem holds for the moment functions. These assumptions are related to Assumption 4.3 in Kaido et al. (2019a), and are also required for our bootstrap procedure to hold. The option (a) or (b) in Assumption 2.3.5(iv) splits the cases when the researcher's objective function depends on W (such as in Example 4) with the cases when the researcher's objective function does not depend on W (such as in subvector inference). Finally, Assumption 2.3.5(v) is the partial identification condition given in Chernozhukov et al. (2007a), equation (4.5), and is useful when establishing the Hausdorff consistency and the rate of convergence of the estimated identified set to the true identified set.

In the following lemma, for any sequence $\{P_n\in\mathcal{P}\}_{n=1}^{\infty}$ converging to the Borel probability measure $P\in\mathcal{P}$, we let $\mathbb{G}_{n,P_n} := \sqrt{n}(\mathbb{P}_n - P_n) \in \ell^\infty(\mathcal{F})$ denote the empirical process indexed by P_n. Adding Assumption 2.3.5, we have the following result:

Lemma 2.3.1. *Suppose Assumptions 2.3.1 - 2.3.5 hold. Then for any sequence* $\{P_n\in\mathcal{P}\}_{n=1}^{\infty} \rightsquigarrow P\in\mathcal{P}$ *we have* $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_P$ *where* $\mathbb{G}_P$ *is a tight Borel measurable element in* $\mathcal{T}_P(\mathcal{F})$, *and:*

$$\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \rightsquigarrow (\Psi_I^{\ell b})'_P(\mathbb{G}_P), \tag{2.37}$$

$$\sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P_n)) \rightsquigarrow (\Psi_I^{ub})'_P(\mathbb{G}_P). \tag{2.38}$$

This result follows from the extended continuous mapping theorem (Theorem 1.11.1 in Van Der Vaart and Wellner (1996)) in combination with the result of Theorem 2.3.1. When combined with Theorem 2.3.1, Lemma 2.3.1 shows that the properly recentered and rescaled value functions converge in distribution to $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$, evaluated at the limiting empirical process $\mathbb{G}_P$, along any converging sequence $\{P_n\in\mathcal{P}\}_{n=1}^{\infty}$ satisfying Assumptions 2.3.1 - 2.3.5. The next section shows that the objects on the right side of (2.37) and (2.38) can be approximated uniformly using a nonparametric bootstrap procedure.

2.3.3 The Bootstrap Version

This section proposes a bootstrap procedure that will allow us to consistently estimate the distributions of the value functions $(\Psi_I^{\ell b}(P), \Psi_I^{ub}(P))$ uniformly over $\mathcal{P}$. In particular, we propose the following approximations:

$$\text{Lower Approximation: } \sqrt{n}\left(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)\right), \tag{2.39}$$

$$\text{Upper Approximation: } \sqrt{n}\left(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)\right). \tag{2.40}$$

We will use the distribution of (2.39) to approximate the distribution of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$, and we will use the distribution of (2.40) to approximate the distribution of $(\Psi_I^{ub})'_P(\mathbb{G}_P)$.

Remark 2.3.3. *Note, unlike typical inference procedures, we do not standardize the moment conditions by their sample standard deviations. However, our procedure is still invariant to rescaling of the moment conditions by the fact that any rescaling will be reflected in the procedure as an equivalent (but opposite) rescaling of the Lagrange multipliers. Furthermore, the imposition of Assumption 2.3.3 allows us to forgo using any moment selection procedure (see Andrews and Soares (2010)) which are typically used in inference problems for partially identified models. This connection between moment selection and constraint qualifications is interesting in its own right.*

We must be precise about the conditions under which the law of the approximations (2.39) and (2.40), conditional on the data $\{W_i\}_{i=1}^n$, can approximate the unconditional law of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$ uniformly over $\mathcal{P}$. Let $\{\{W_i^b\}_{i=1}^n : b = 1, \ldots, B\}$ denote the bootstrap samples. We maintain the following assumption:

Assumption 2.3.6. *The bootstrap samples $\{W_i^b\}_{i=1}^n$ for $b = 1, \ldots, B$, are drawn i.i.d. with replacement from the original sample $\{W_i\}_{i=1}^n$.*

The following lemma, which is necessary for our main result, shows that the proposed bootstrap procedure is uniformly valid:

Lemma 2.3.2. *Suppose that conditional on $\{W_i\}_{i=1}^n$ we have that, uniformly over $\mathcal{P}$, $\mathbb{G}_n^b \rightsquigarrow \mathbb{G}_P$ where $\mathbb{G}_P$ is a tight random element in $\ell^\infty(\mathcal{F})$. Then under Assumptions 2.3.1-2.3.6:*

$$\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n))|\{W_i\}_{i=1}^n \rightsquigarrow (\Psi_I^{\ell b})'_P(\mathbb{G}_P),$$

$$\sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n))|\{W_i\}_{i=1}^n \rightsquigarrow (\Psi_I^{ub})'_P(\mathbb{G}_P).$$

A confidence set for the true parameter ψ_0 can then be constructed using the quantiles of the bootstrapped distributions of (2.39) and (2.40). In particular, the confidence set $C_n^\psi(1-\alpha)$ with asymptotic coverage probability of $1-\alpha$ can be constructed as:

$$C_n^\psi(1-\alpha) := \left[\Psi_I^{\ell b}(\mathbb{P}_n) - \frac{\hat{\Psi}_\alpha^{\ell b}}{\sqrt{n}}, \Psi_I^{ub}(\mathbb{P}_n) + \frac{\hat{\Psi}_\alpha^{ub}}{\sqrt{n}}\right], \tag{2.41}$$

where the pair $(\hat{\Psi}_\alpha^{\ell b}, \hat{\Psi}_\alpha^{ub})$ minimize the length of the confidence set $C_n^\psi(1-\alpha)$ subject to the constraints:

$$P_n^b\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)) + \sqrt{n}\Delta(\mathbb{P}_n)\right) \geq 1-\alpha, \tag{2.42}$$

$$P_n^b\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)) - \sqrt{n}\Delta(\mathbb{P}_n) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n))\right) \geq 1-\alpha, \tag{2.43}$$

where P_n^b is the bootstrap distribution and Δ is the length of the identified set. Note that under Assumption 2.3.3, we will rule out cases where length of the identified set can be drifting towards zero and thus we avoid issues of uniformity that occur in this scenario (see Stoye (2009b)).

The following result verifies that under our assumptions, the confidence set given in (2.41) is uniformly asymptotically valid:

Theorem 2.3.2. *Under Assumptions 2.3.1 - 2.3.6,*

$$\liminf_{n\to\infty} \inf_{\{(\psi,P):\ \psi\in\Psi_I(P),\ P\in\mathcal{P}\}} P\left(\psi \in C_n^{\psi}(1-\alpha)\right) \geq 1-\alpha, \tag{2.44}$$

where $C_n^{\psi}(1-\alpha)$ *is as in* (2.41).

The confidence set $C_n^{\psi}(1-\alpha)$ is both conceptually simple and easy to implement. Indeed, computing the confidence set amounts to bootstrapping the value functions for the optimization problems that define the endpoints of the set $\Psi_I(\cdot)$. Calibrating the critical values $\hat{\Psi}_{\alpha}^{\ell b}$ and $\hat{\Psi}_{\alpha}^{ub}$ is then easily done once the bootstrap distribution has been recovered. In other words, Assumptions 2.3.1 - 2.3.6 are sufficient for a researcher to "naively" bootstrap the value functions of a linear program in order to construct a uniformly valid confidence set for a linear functional of interest.

Intuitively, most of the "heavy lifting" required to prove Theorem 2.3.2 has already been completed in the proofs of Lemmas 2.3.1 and Lemma 2.3.2, and both of these Lemmas rely crucially on Theorem 2.3.1. Most of the assumptions needed to obtain Theorem 2.3.2 are analogous to assumptions made previously in the literature in partial identification (c.f. Bugni et al. (2015), Bugni et al. (2017) and Kaido et al. (2019a)), the important exception being Assumption 2.3.3. Indeed, the simplicity of our procedure relative to previous approaches might aptly be characterized as arising primarily from Assumption 2.3.3, which is motivated by analogous assumptions in the literature on optimization. However, as noted by Kaido et al. (2019b), even Assumption 2.3.3 can be recognized in various forms in the literature in partial identification. In the next section, we will attempt to provide the reader with some further intuition regarding Assumption 2.3.3.

2.4 Further Discussion

This section provides some additional discussion of the method proposed in the previous section. In particular, this section will attempt to provide some further intuition for Assumption 2.3.3, and will then discuss the case when the identified set is empty in finite sample.

2.4.1 Constraint Qualifications and Uniqueness of Lagrange Multipliers

Researchers may be concerned about imposing a uniform version of the LICQ, as is implied by Assumption 2.3.3(ii). Indeed, this is a somewhat non-standard assumption in the econometrics literature, although various forms of constraint qualifications appear in many papers on subvector inference in partial identification (see Kaido et al. (2019b)). In this section we show that, at least for a fixed data-generating measure P, the cases in which the LICQ is not satisfied are somewhat pathological, or "non-generic," in a sense to be made precise shortly.[5]

To state the result, let us suppose for simplicity that the researcher has only moment inequality constraints, and let $M_P(\theta)$ denote the column vector with rows $\{Pm_j(W,\theta)\}_{j=1}^k$. For any feasible θ we must have $M_P(\theta) \leq \mathbf{0}$. We now consider a $\epsilon-$perturbation of the moment conditions, so that the perturbed model is satisfied when $M_P(\theta) \leq \epsilon$, where $\epsilon := (\epsilon_1, \ldots, \epsilon_k)^T \in \mathcal{E}$ is a perturbation parameter. We will take $\mathcal{E} = \mathbb{R}^k$, and we will equip $\mathcal{E}$ with a probability measure P_ϵ that is absolutely continuous with respect to the Lebesgue measure. Finally, let $\boldsymbol{G}_\epsilon(\theta, P)$ be the matrix with rows $\{\nabla_\theta Pm_j(W,\theta)\}_{j\in\mathscr{A}_\epsilon(\theta,P)}$, where $\mathscr{A}_\epsilon(\theta,P)$ is an index set for the binding moment inequalities at (θ, P) with perturbation ϵ; i.e., the moment inequalities that satisfy $Pm_j(W,\theta) = \epsilon_j$ at (θ, P).

We can now present an interesting proposition, which is derived from a result in differential topology. To state and prove the result, recall that a point θ is called a *critical point* of a map $f : \Theta \to \mathbb{R}^{d_f}$ if the Jacobian

[5] We are grateful to Victor Aguirregabiria for this suggestion.

$\nabla_\theta f(\theta)$ does not have full row rank at θ. For any such θ, the value $y = f(\theta)$ is called a *critical value*. Sard's Theorem from differential topology then says that if f is sufficiently smooth, then the set of critical values for f has Lebesgue measure zero in $\mathbb{R}^{d_f}$. Using this result, we can obtain the following proposition due to Spingarn and Rockafellar (1979):

Proposition 2.4.1. *Suppose that $M_P(\theta)$ is $r-$times continuously differentiable, where $r \geq \max\{1, d_\theta - k + 1\}$. Then $P_\epsilon-$almost surely, for any θ satisfying $M_P(\theta) \leq \epsilon$ we will have:*

$$eig\left(\boldsymbol{G}_\epsilon(\theta, P)\boldsymbol{G}_\epsilon(\theta, P)^T\right) > 0. \tag{2.45}$$

Proof. Up to a change in notation, the proof follows *exactly* from the proof of Spingarn and Rockafellar (1979) Theorem 1, and is included only for completeness. Fix any θ satisfying $M_P(\theta) \leq \epsilon$. Let $\mathcal{I} = \{1, \ldots, k\}$, and let $A \subset \mathcal{I}$ be any subset. Denote by $M_P^A(\theta) : \Theta \to \mathbb{R}^{|A|}$ the subvector of $M_P(\theta)$ that contains the elements of $M_P(\theta)$ indexed by A. Then by Sard's Theorem we have that the set of critical values for $M_P^A(\theta)$ have measure zero in $\mathbb{R}^{|A|}$. Denoting by ϵ^A the projection of ϵ onto $\mathbb{R}^{|A|}$, we have that:

$$N(A) := \{\epsilon \in \mathbb{R}^k : \epsilon^A \text{ is a critical value for } M_P^A(\theta)\},$$

has measure zero under P_ε. Repeating the exercise for every $A \subseteq \mathcal{I}$ we have:

$$N := \bigcup_{A \subseteq \mathcal{I}} N(A),$$

has measure zero under P_ε. Thus $\mathcal{E} \setminus N$ has probability 1. Now take any $\epsilon \in \mathcal{E} \setminus N$, and take $A = \mathscr{A}_\epsilon(\theta, P)$. Then the rows of $\boldsymbol{G}_\epsilon(\theta, P)$—formed by the gradients $\{\nabla_\theta P m_j(W, \theta)\}_{j \in \mathscr{A}_\epsilon(\theta, P)}$—are linearly independent. This completes the proof. ■

This result shows that even if the initial moment conditions do not satisfy the LICQ implied by Assumption 2.3.3(ii), if we perturb the moment conditions slightly, then at any feasible value for the perturbed conditions the LICQ will hold with probability 1. This illustrates that cases where Assumption 2.3.3 fails are truly "knife-edge" cases. In the optimization literature, these results are referred to as genericity results, since they show that "generic" (or $P_\epsilon-$almost all) convex programs satisfy properties like Assumption 2.3.3. A similar analysis can be repeated for the case with both equality and inequality constraints by first converting all equality constraints into two paired inequality constraints, and then choosing the support of the perturbation parameter in a way to ensure that the two paired inequality constraints are "separated" with probability 1.

An important caveat is that this analysis holds in the case when the probability measure P is fixed. Indeed, we have been unable to construct an analogous perturbation analysis that can be used to justify the LICQ uniformly over $\mathcal{P}$, although we feel this will be a fruitful avenue for future research in partial identification. Regardless, we feel that Proposition 2.4.1 helps to put the LICQ Assumption in perspective.

2.4.2 Empty Sets

In some cases the estimated identified set may be empty in finite samples even though the true DGP satisfies the assumptions in this chapter. However, when the identified set is empty in finite sample it is possible to "relax" the moment conditions to the point where the relaxed moment conditions have nonempty interior. We might then perform our subvector inference procedure on this relaxed version of the identified set. If the model is correctly specified, then this "relaxation" of the moment conditions can gradually be lifted. We will summarize this procedure here.

Consider the following relaxed versions of the convex programs (2.29) and (2.30):

$$\Psi_I^{\ell b}(\mathbb{P}_n, c_n) := \inf_{\theta \in \Theta} \quad \mathbb{P}_n \psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq c_n, \qquad j = 1, \ldots, k,$$

$$\Psi_I^{ub}(\mathbb{P}_n, c_n) := \sup_{\theta \in \Theta} \quad \mathbb{P}_n \psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq c_n, \qquad j = 1, \ldots, k.$$

For convenience, we will take the infimum over the empty set to be $+\infty$ and the supremum over the empty set to be $-\infty$. Now define:[6]

$$c_n^* := \inf \left\{ c_n \in [0, +\infty) : \Psi_I^{\ell b}(\mathbb{P}_n, c_n) < +\infty \right\}.$$

Then by definition the following programs have nonempty feasible sets:

$$\Psi_I^{\ell b}(\mathbb{P}_n, c_n^*) := \inf_{\theta \in \Theta} \quad \mathbb{P}_n \psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq c_n^* + \varepsilon, \qquad j = 1, \ldots, k, \tag{2.46}$$

$$\Psi_I^{ub}(\mathbb{P}_n, c_n^*) := \sup_{\theta \in \Theta} \quad \mathbb{P}_n \psi(W, \theta) \qquad \text{s.t.} \qquad \mathbb{P}_n m_j(W, \theta) \leq c_n^* + \varepsilon, \qquad j = 1, \ldots, k, \tag{2.47}$$

where the extra ε ensures that the moment inequalities have nonempty interior, which is necessary although not sufficient for Assumption 2.3.3(ii) to hold (see the discussion following Assumption 2.3.3). Now note that if $c_n^* = o(n^{-1/2})$, and if Assumptions 2.3.1 - 2.3.6 are satisfied, then the value functions $\Psi_I^{\ell b}(\mathbb{P}_n, c_n^*)$ and $\Psi_I^{ub}(\mathbb{P}_n, c_n^*)$ from (2.46) and (2.47) can be used in place of the value functions $\Psi_I^{\ell b}(\mathbb{P}_n)$ and $\Psi_I^{ub}(\mathbb{P}_n)$ from (2.27) and (2.28).

This procedure is very similar to the idea of a "misspecification-robust identified set" recently introduced by Andrews and Kwon (2019). Indeed, the relaxation parameter c_n^* guarantees under our Assumptions that the identified set for ψ_0 will always be non-empty. Different variations on the notion of a "misspecification-robust identified set" are also possible.[7] If the relaxation parameter satisfies $c_n^* = o(n^{-1/2})$, then our procedure remains a valid inference procedure for ψ_0; if not, then the model is misspecified, but our procedure remains valid for a "pseudo-true" value of ψ_0 defined by the relaxed moment conditions. We refer readers to Andrews and Kwon (2019) for a further discussion of this idea.

2.5 Simulation Evidence

To practically test the proposed the procedure, we performed Monte Carlo experiments on three different economic examples. In particular, we consider two canonical partial identification examples—given by the missing data problem from Example 1 and the linear regression example with interval-valued dependent variable from Example 2—as well as a less canonical example, given by the problem of inference on counterfactual policies in Example 4. For brevity in the main text, we have placed the missing data example, and the interval-valued regression example in Appendix 2.B, and will only describe the DGP and results for Example 4 here. However, as Appendix 2.B shows, the inference procedure also performed well in the missing data and the interval-valued regression examples.

[6] Equivalently: $c_n^* := \inf \left\{ c_n \in [0, +\infty) : \Psi_I^{ub}(\mathbb{P}_n, c_n) > -\infty \right\}$.

[7] Indeed, our notion here differs from Andrews and Kwon (2019) in the sense that the choice of c_n^* is more conservative, but computationally simpler than the relaxation proposed in Andrews and Kwon (2019).

2.5.1 Description

Recall Example 4 from Kasy (2016) on inference on counterfactual treatment policies. In that example, we had $g_d(X) := \mathbb{E}[Y_d|X]$, which was assumed to be obtained from an initial study on the effect of some treatment D, but is only partially identified and known to lie in the (estimated) set $\mathcal{G}_d$. The policy maker now wants to determine the effect of a treatment policy in a new population with distribution of covariates given by $P(X = x)$.

The policy maker compares two policies, A and B, which are defined by the conditional probability of being assigned to treatment given $X = x$. In particular, policy A is associated with the conditional treatment assignment probability $P(D^A = 1|X = x)$ and policy B is associated with the conditional treatment assignment probability given by $P(D^B = 1|X = x)$. Let $h^{AB}(x) = P(D^A = 1|X = x) - P(D^B = 1|X = x)$. Furthermore, let $\psi^{AB} = \mathbb{E}[Y^A - Y^B]$, that is, the expected difference in outcomes under policies A and B. Then the identified set for ψ^{AB} is given by:

$$\Psi_{lb}^{AB}(P) = \inf_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} h^{AB}(x)\,(g_1(x) - g_0(x))\,P(X = x), \tag{2.48}$$

$$\Psi_{ub}^{AB}(P) = \sup_{(g_0,g_1)\in\mathcal{G}_0\times\mathcal{G}_1} \sum_{x\in\mathcal{X}} h^{AB}(x)\,(g_1(x) - g_0(x))\,P(X = x). \tag{2.49}$$

To motivate the relevance and guide the construction of our simulation study, we can consider the case study in Kasy (2016), in which the initial study to determine bounds on $g_d(\cdot)$ was the Tennessee Star experiment. The Tennessee Star experiment saw students randomized within schools to small and large classrooms. The outcome in this experiment was student performance on standardized tests, in particular for reading and math. While student assignment to small and large class sizes was random, compliance was imperfect for a number of reasons. The study includes a variety of covariates, including indicators for whether the student was female, black, or was enrolled to receive a free lunch (an indicator for poverty).

As in the Tennessee STAR experiment case study, we will consider a data generating process that includes a binary instrument Z, a treatment variable D, potential outcomes Y_0 and Y_1, and a vector of covariates $X = (X^a, X^b)$. Since we will take both X^a and X^b as binary, it will be equivalent (and more notationally beneficial) to see X as a scalar covariate that takes values in the set $\{x_1, x_2, x_3, x_4\}$. The instrument is generated to satisfy $Z \perp\!\!\!\perp (Y_0, Y_1)$, and the DGP for the instrument and treatment variable is given by:

$$\begin{aligned} &Z \sim \text{Bernoulli}(0.5) \\ &D = \mathbb{1}\{(2Z-1) > \max\{c/\sqrt{n}, \delta\} \cdot \varepsilon\}, \quad \varepsilon \sim N(0,1), \text{ and } c \in \{0,1,2\}. \end{aligned}$$

We will allow $c \in \{1, 10, 20\}$ to vary across DGPs. While the precise values of c are chosen arbitrarily, varying the values of c can be used to control the size of the identified set by changing the relationship between D and Z. Indeed, if $c = 0$ and $\delta = 0$, we have $D = Z$, and the conditional average treatment effect will be point-identified. On the other hand, if c is very large then the dependence between Z and D is weak, and the identified set for the conditional average treatment effect will be large. In the DGP we fix $\delta = 10^{-6}$ in order to ensure the model is always partially identified, even as $n \to \infty$.

We restrict our outcome variable Y to be in the range $\mathcal{Y} := \{1, 2, 3, 4, 5\}$. Returning to the Tennessee STAR experiment example, this might correspond to quintiles of the standardized test distribution, or some other mapping from test scores. The conditional distribution $Y_0|X$ is specified as follows:

$$\begin{aligned} Y_0|X = x_1 &\sim \text{Categorical}(5, p_1 = 0.2,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.2,\ p_5 = 0.2), \\ Y_0|X = x_2 &\sim \text{Categorical}(5, p_1 = 0.3,\ p_2 = 0.25,\ p_3 = 0.25,\ p_4 = 0.1,\ p_5 = 0.1), \end{aligned}$$

$$Y_0|X = x_3 \sim \text{Categorical}(5, p_1 = 0.3,\ p_2 = 0.25,\ p_3 = 0.25,\ p_4 = 0.1,\ p_5 = 0.1),$$
$$Y_0|X = x_4 \sim \text{Categorical}(5, p_1 = 0.4,\ p_2 = 0.35,\ p_3 = 0.25,\ p_4 = 0,\ p_5 = 0).$$

Treating the values in $\mathcal{Y}$ as analogous to test scores, we can say a few things about this DGP. First, individuals with $X = x_1$ are equally likely to obtain any test score in the range $\mathcal{Y}$. However, individuals with $X = x_2$ or $X = x_3$ have the same test distribution, and are more likely than individuals with $X = x_1$ to obtain a lower score. Finally, individuals with $X = x_4$ are more likely to obtain a worse test score than any other subpopulation. With this in mind, the conditional distribution $Y_1|X$ is specified as follows:

$$Y_1|X = x_1 \sim \text{Categorical}(5, p_1 = 0.2,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.2,\ p_5 = 0.2),$$
$$Y_1|X = x_2 \sim \text{Categorical}(5, p_1 = 0.1,\ p_2 = 0.1,\ p_3 = 0.25,\ p_4 = 0.25,\ p_5 = 0.3),$$
$$Y_1|X = x_3 \sim \text{Categorical}(5, p_1 = 0.1,\ p_2 = 0.1,\ p_3 = 0.25,\ p_4 = 0.25,\ p_5 = 0.3),$$
$$Y_1|X = x_4 \sim \text{Categorical}(5, p_1 = 0,\ p_2 = 0,\ p_3 = 0.25,\ p_4 = 0.35,\ p_5 = 0.4).$$

Note that when specifying the conditional distribution for $Y_1|X$, we have simply reversed the order of the probabilities from the conditional distribution for $Y_0|X$; in particular, if the probability vector parameterizing the categorical distribution for $Y_0|X = x$ was $p = (p_1, p_2, p_3, p_4, p_5)^T$, then the probability vector parameterizing the categorical distribution for $Y_1|X = x$ is is given by $p' = (p_5, p_4, p_3, p_2, p_1)^T$. The implications of this DGP is that, on average, we do not expect individuals with $X = x_1$ to lose or gain from treatment in terms of improved test scores, whereas individuals in the populations $X = x_2, x_3$ or x_4 will all see improved test scores from treatment (with those with $X = x_4$ benefiting the most on average). We assume that the initial and target population distribution of covariates is given by:

$$\begin{aligned}
\text{Initial Distribution: } & P(X = x_1) = 0.25, & \text{Target Distribution: } & P(X = x_1) = 0.3,\\
& P(X = x_2) = 0.25, & & P(X = x_2) = 0.3,\\
& P(X = x_3) = 0.25, & & P(X = x_3) = 0.2,\\
& P(X = x_4) = 0.25, & & P(X = x_4) = 0.2.
\end{aligned}$$

Note that in our simulations we will take draws from the target distribution, so that the policy-maker will have sampling uncertainty arising from lack of perfect knowledge of the population covariate proportions. Finally, in our setup the policy-maker compares the policies A and B represented by the following treatment assignment rules:

$$\begin{aligned}
& P(D^A = 1|X = x_1) = 0, & & P(D^B = 1|X = x_1) = 0.5,\\
& P(D^A = 1|X = x_2) = 0.75, & & P(D^B = 1|X = x_2) = 0.5,\\
& P(D^A = 1|X = x_3) = 0.75, & & P(D^B = 1|X = x_3) = 0.5,\\
& P(D^A = 1|X = x_4) = 1, & & P(D^B = 1|X = x_4) = 0.5.
\end{aligned}$$

In other words, policy A gives a highly unequal treatment assignment probability across covariate values, whereas policy B represents a policy that is more egalitarian in the sense that the treatment assignment probability does not depend on the covariate values. However, note that policy A assigns the highest treatment assignment probability to individuals who are in covariate groups with the highest conditional average treatment effect. In contrast, policy B assigns equal treatment assignment probability to all groups, including the group represented by $X = x_1$, which has zero conditional average treatment effect. Quick computation shows that policy A, which assigns more weight to those who benefit from treatment, will be

preferred to policy B.

For reference, according to this DGP the population values of the parameters of interest are given by the following:

$$\begin{aligned} &\mathbb{E}[Y_1 - Y_0|X = x_1] = 0, && \psi^{AB} = 0.3675, \\ &\mathbb{E}[Y_1 - Y_0|X = x_2] = 1.1, \\ &\mathbb{E}[Y_1 - Y_0|X = x_3] = 1.1, \\ &\mathbb{E}[Y_1 - Y_0|X = x_4] = 2.3. \end{aligned}$$

The moment conditions used to bound the conditional average treatment effect come from Russell (2019). In particular, for each fixed $X = x$ the sharp set of constraints on the conditional distribution $P(Y_0 = y_0, Y_1 = y_1, D = d|X = x, Z = z)$ when $Y_0, Y_1, X \perp\!\!\!\perp Z$ are derived in Russell (2019). For the sake of brevity we will not discuss these constraints in detail here, although we will note that all the constraints are linear and can be expressed in terms of the distribution of the observable variables (Y, D, X, Z) only. Letting $\mathcal{P}(x, z)$ denote the sharp set of all conditional distributions $p_{y_0,y_1,d}(x, z) := P(Y_0 = y_0, Y_1 = y_1, D = d|X = x, Z = z)$ satisfying $Y_0, Y_1, X \perp\!\!\!\perp Z$, and noticing that:

$$g_1(x) - g_0(x) := \mathbb{E}[Y_1 - Y_0|X = x] = \sum_{y_0,y_1,d,z} (y_1 - y_0) p_{y_0,y_1,d}(x, z) P(Z = z),$$

we obtain:

$$\Psi^{AB}_{\ell b}(P) = \inf_{p_{y_0,y_1,d}(x,z)\in\mathcal{P}(x,z)} \sum_{x\in\mathcal{X}} h^{AB}(x) \left(\sum_{y_0,y_1,d,z} (y_1 - y_0) p_{y_0,y_1,d}(x, z) P(Z = z) \right) P(X = x), \tag{2.50}$$

$$\Psi^{AB}_{ub}(P) = \sup_{p_{y_0,y_1,d}(x,z)\in\mathcal{P}(x,z)} \sum_{x\in\mathcal{X}} h^{AB}(x) \left(\sum_{y_0,y_1,d,z} (y_1 - y_0) p_{y_0,y_1,d}(x, z) P(Z = z) \right) P(X = x), \tag{2.51}$$

which are also linear programs. The partially identified parameter vector $p_{y_0,y_1,d}(x, z)$ contains 50 elements for each fixed $X = x$ and $Z = z$, so that in total the partially identified parameter vector has 400 elements. While we recognize that there are likely simpler ways of constructing this bounding problem, the larger dimension of the partially identified parameter vector serves as a useful illustration of the computational benefit of our approach.

In all Monte Carlo exercises we take $B = 1000$ bootstrap samples for each experiment, and we implement each experiment 1000 times to determine the simulated coverage probability. We also consider various sample sizes $n \in \{100, 250, 500, 1000\}$. In each DGP, we also threshold the length of the identified set; i.e. we use $\Delta_n^* = \mathbb{1}\{\Delta_n > b_n\}$, with $b_n = (\log(n))^{-1/2}$, rather than Δ_n when computing the critical values from (2.42) and (2.43). We find this thresholding helps to improve the coverage in finite sample in cases when the model is close to point identification, and introduces at most a conservative distortion under the assumptions in this chapter.

2.5.2 Results

The simulation results for the interval valued regression example are displayed in Table 2.1. Similar to the Monte Carlo excercises for the missing data and interval-valued regressor examples in Appendix 2.B, the coverage probability for the true parameter is slightly above nominal in most of the DGPs considered. This results from the fact that often the true parameter lies interior to the identified set, as well as from the

thresholding discussed at the beginning of this section. As the value of c increases, we see that the length of the identified set increases since the effect of the instrument on selection becomes weaker. However, for most of the DGPs considered our confidence set remain informative; in particular, given a reasonably small sample size (~ 500), the policy-maker is always able to conclude that policy A is significantly better than policy B for any value of c we considered. This problem is slightly more computationally involved than both the missing data example and interval valued regression considered in Appendix 2.B, but we still find that the approximate time to compute a confidence set is only around 10 seconds.[8] Compared to other procedures this is extremely fast, especially since our partially identified parameter vector $p_{y_0,y_1,d}(x,z)$ has 400 elements.

2.6 Conclusion

This chapter proposes a simple procedure for constructing confidence intervals for functionals of a partially identified parameter vector. The procedure approximates the distribution of the upper and lower bounds of the identified set for the functional of interest through a simple bootstrap procedure. In particular, we show that if the problem is sufficiently regular, a "naive" bootstrap procedure can be used, where the researcher (essentially) repeatedly solves a linear program, and computes confidence sets by taking appropriate quantiles of the bootstrap distribution of the value functions. Uniform validity of this "naive" procedure is proven by making connections to results in the Operations Research literature on stochastic programming, and in particular by appealing to the notion of uniform Hadamard directional differentiability. The procedure is found to be extremely computationally efficient, even when the parameter vector is very high-dimensional; indeed, the parameter vector had 400 elements in the simulation exercise presented in the main text, and a confidence set for a linear functional could still be constructed in about 10 seconds. The most important condition for the validity of our procedure is found to be the existence and uniqueness of optimal solutions and Lagrange multipliers, and we feel that the development of more primitive conditions to ensure these conditions hold will be a research project worthy of further pursuit.

[8] All Monte Carlo exercises were run on a laptop computer with an Intel Core i7-8550U CPU.

Appendix 2.A Proofs

Throughout this appendix we use the following notation: if X_n, X are maps in a metric space $(\mathbb{D}, d)$ then:

- $X_n = o_{\mathcal{P}}(a_n)$ is used to denote uniform (over $\mathcal{P}$) convergence in probability of the random element $|X_n/a_n|$ to 0; i.e. $\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P_P^*(|X_n/a_n| > \varepsilon) = 0$ for every $\varepsilon > 0$,
- $X_n = O_{\mathcal{P}}(a_n)$ is used to denote uniform (over $\mathcal{P}$) stochastic boundedness of the random element $|X_n/a_n|$; i.e. the fact that for any $\varepsilon > 0$ there exists a a finite M and an N such that $\sup_{P\in\mathcal{P}} P_P^* (|X_n/a_n| > M) < \varepsilon$ for all $n \geq N$.

We will also rely on the following facts which are not proven here, but for which references are provided.

Fact 2.A.1. *Suppose that $\{P_n \in \mathcal{P}_+\}_{n=1}^{\infty} \rightsquigarrow P \in \mathcal{P}$. Under Assumption 2.3.1, 2.3.2 and 2.3.3 there exists an N such that for all $n \geq N$ strong duality holds for $P_n \in \mathcal{P}_+$; that is, if $\mathcal{L}(\theta, \lambda)(P_n)$ is the Lagrangian at probability measure P, then*

$$\Psi_I^{\ell b}(P_n) = \inf_{\theta\in\Theta} \sup_{\lambda\geq 0} \mathcal{L}(\theta, \lambda)(P_n) = \sup_{\lambda\geq 0} \inf_{\theta\in\Theta} \mathcal{L}(\theta, \lambda)(P_n),$$

and

$$\Psi_I^{ub}(P_n) = \sup_{\theta\in\Theta} \inf_{\lambda\leq 0} \mathcal{L}(\theta, \lambda)(P_n) = \inf_{\lambda\leq 0} \sup_{\theta\in\Theta} \mathcal{L}(\theta, \lambda)(P_n).$$

This result is called Lagrangian Duality in convex optimization; see, for example, Borwein and Lewis (2010) Theorem 4.3.7. This follows since any sequence $\{P_n \in \mathcal{P}_+\}_{n=1}^{\infty} \rightsquigarrow P \in \mathcal{P}$ must eventually lie in $\mathcal{P}^{\varepsilon}$, so that Assumption 2.3.3 holds in the tails of any such sequence.

Before the next fact, some definitions:

Definition 2.A.1 (Upper Hemicontinuity)**.** *For metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a correspondence $G : \mathcal{X} \to \mathcal{Y}$ is said to be upper hemicontinuous at $x \in \mathcal{X}$ if for every open subset S of $\mathcal{Y}$ with $G(x) \subseteq S$ there exists a $\delta > 0$ such that $G(B_\delta(x)) \subseteq S$.*

Definition 2.A.2 (Compact-Valued)**.** *For metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a correspondence $G : \mathcal{X} \to \mathcal{Y}$ is said to be compact-valued if $G(x)$ is a compact subset of $\mathcal{Y}$ for each $x \in \mathcal{X}$.*

Definition 2.A.3 (Closed at x)**.** *For metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a correspondence $G : \mathcal{X} \to \mathcal{Y}$ is said to be closed at x if for any sequence $\{x_n\}$ and $\{y_n\}$ with $x_n \to x$ and $y_n \to y$ we have that $y \in G(x)$ whenever $y_n \in G(x_n)$ for all n.*

Fact 2.A.2 (Proposition E.2 in Ok (2007))**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces and $\Gamma : \mathcal{X} \to \mathcal{Y}$ a correspondence. If Γ is compact-valued and upper hemicontinuous at $x \in \mathcal{X}$, then for any sequence $\{x_m\}_{m=1}^{\infty} \subset \mathcal{X}$ and $\{y_m\}_{m=1}^{\infty} \subset \mathcal{Y}$ with $x_m \to x$ and $y_m \in \Gamma(x_m)$ for each m, there exists a subsequence $\{y_{m_k}\}_{k=1}^{\infty}$ such that $y_{m_k} \to y \in \Gamma(x)$.*

Finally, $\Theta_{\ell b}(P)$ and $\Theta_{ub}(P)$ denote the set of optimal solutions to (2.27) and (2.28), and $\Lambda_{\ell b}(P)$ and $\Lambda_{ub}(P)$ denote the set of Lagrange multipliers for (2.27) and (2.28).

2.A.1 Proof of Results in Main Text

Remark 2.A.1. *The following proof follows similar steps to the proof of Theorem 7.24 in Shapiro et al. (2009), which shows Hadamard directional differentiability. However, the proof here establishes that this*

property holds "uniformly" over $\mathcal{P}$ under the assumptions of Theorem 2.3.1. The proof of uniformity follows namely from (i) the assumption $h_n \to h$ in the sup norm (and thus uniformly) where h is an operator that is uniformly continuous with respect to θ (the latter is provided by Lemma 2.A.5), (ii) boundedness of the Lagrangian (given by Lemma 2.A.3) and (iii) continuity of the optimal solutions and Lagrange multipliers (given by Lemma 2.A.2).

Proof of Theorem 2.3.1. We can focus on the lower bound, since the upper bound can be treated analogously. Consider any converging sequences $P_n \rightsquigarrow P \in \mathcal{P}$, $\{h_n\} \subset \ell^\infty(\mathcal{F})$ and $\{t_n\} \subset \mathbb{R}_+$ with $t_n \downarrow 0$ and $h_n \to h \in \mathcal{T}_P(\mathcal{F})$ such that $P_n + t_n h_n \in \mathcal{P}_+$ for all $n \geq 1$. Recall the Lagrangian at a probability measure P_n is given by:

$$\mathcal{L}(\theta, \lambda)(P_n) := P_n \psi(W, \theta) + \sum_{j=1}^{k} \lambda_j P_n m_j(W, \theta), \tag{2.52}$$

where $\lambda := (\lambda_1, \ldots, \lambda_k)' \in \overline{\mathbb{R}}_+^k$ is a vector of Lagrange multipliers. Denote the "unperturbed" and "perturbed" programs respectively as:

$$\Psi_I^{\ell b}(P_n) := \inf_{\theta \in \Theta} \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}(\theta, \lambda)(P_n), \tag{2.53}$$

$$\Psi_I^{\ell b}(P_n + t_n h_n) := \inf_{\theta \in \Theta} \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}(\theta, \lambda)(P_n + t_n h_n), \tag{2.54}$$

where $P_n + t_n h_n$ is interpreted elementwise. By Fact 2.A.1 we have by Lemma 2.A.1 that there exists an N such that for all $n \geq N$:

$$\Psi_I^{\ell b}(P_n) = \inf_{\theta \in \Theta} \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}(\theta, \lambda)(P_n) = \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \inf_{\theta \in \Theta} \mathcal{L}(\theta, \lambda)(P_n), \tag{2.55}$$

$$\Psi_I^{\ell b}(P_n + t_n h_n) = \inf_{\theta \in \Theta} \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}(\theta, \lambda)(P_n + t_n h_n) = \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \inf_{\theta \in \Theta} \mathcal{L}(\theta, \lambda)(P_n + t_n h_n). \tag{2.56}$$

Lemma 2.A.1 implies there exists an optimal $\theta_{\ell b}^*(P_n)$ for each $n \geq N$. Now consider the sequence $\{\theta_{\ell b}^*(P_n)\}_{n=1}^\infty$ with $\theta_{\ell b}^*(P_n)$ optimal for each $n \geq N$, and conclude that for all $n \geq N$:

$$\Psi_I^{\ell b}(P_n) = \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}\left(\theta_{\ell b}^*(P_n), \lambda\right)(P_n), \tag{2.57}$$

$$\Psi_I^{\ell b}(P_n + t_n h_n) \leq \sup_{\lambda \in \overline{\mathbb{R}}_+^k} \mathcal{L}(\theta_{\ell b}^*(P_n), \lambda)(P_n + t_n h_n), \tag{2.58}$$

where (2.57) follows from strong duality, and (2.58) follows from the fact that $\theta_{\ell b}^*(P_n)$ is optimal for program (2.53) but not necessarily program (2.54).

By Lemma 2.A.1, we have that there exists a optimal vector of Lagrange multipliers in (2.58) for $n \geq N$. Let $\{\lambda_{\ell b}^*(P_n + t_n h_n)\}_{n=1}^\infty$ be a sequence with $\lambda_{\ell b}^*(P_n + t_n h_n)$ optimal for each $n \geq N$. For any such sequence, note from (2.57) and (2.58) we have for all $n \geq N$:

$$\Psi_I^{\ell b}(P_n) \geq \mathcal{L}\left(\theta_{\ell b}^*(P_n), \lambda_{\ell b}^*(P_n + t_n h_n)\right)(P_n), \tag{2.59}$$

$$\Psi_I^{\ell b}(P_n + t_n h_n) \leq \mathcal{L}\left(\theta^*_{\ell b}(P_n), \lambda^*_{\ell b}\left(P_n + t_n h_n\right)\right)\left(P_n + t_n h_n\right). \tag{2.60}$$

Finally, also note that since $h_n \to h \in \mathcal{T}_P(\mathcal{F})$ by assumption, we have that:

$$h_n = h + o(1). \tag{2.61}$$

Thus, for all $n \geq N$:

$$\begin{aligned}
&\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n) \\
&\leq \mathcal{L}\left(\theta^*_{\ell b}(P_n), \lambda^*_{\ell b}\left(P_n + t_n h_n\right)\right)\left(P_n + t_n h_n\right) - \mathcal{L}\left(\theta^*_{\ell b}(P_n), \lambda^*_{\ell b}(P_n + t_n h_n)\right)(P_n) \quad \text{(from (2.59) and (2.60))} \\
&= t_n h_{n,1} \psi(W, \theta^*_{\ell b}(P_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b, j}\left(P_n + t_n h_n\right) t_n h_{n,j+1} m_j(W, \theta^*_{\ell b}(P_n)) \quad \text{(by (2.52))} \\
&= t_n \left(h_1 \psi(W, \theta^*_{\ell b}(P_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b, j}\left(P_n + t_n h_n\right) h_{j+1} m_j(W, \theta^*_{\ell b}(P_n)) \right) + o(t_n), \quad \text{(by (2.61))}
\end{aligned}$$

where the final line follows from uniform boundedness of the Lagrangian from Lemma 2.A.3(ii). Thus for any sequence $\{\theta^*_{\ell b}(P_n)\}$:

$$\begin{aligned}
&\limsup_{n \to \infty} \frac{\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n)}{t_n} \\
&\leq \limsup_{n \to \infty} \left[h_1 \psi(W, \theta^*_{\ell b}(P_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b, j}\left(P_n + t_n h_n\right) h_{j+1} m_j(W, \theta^*_{\ell b}(P_n)) \right] \\
&= h_1 \psi(W, \theta^*_{\ell b}(P)) + \sum_{j=1}^{k} \lambda^*_{\ell b, j}\left(P\right) h_{j+1} m_j(W, \theta^*_{\ell b}(P)).
\end{aligned} \tag{2.62}$$

The last line follows by convergence of $P_n \rightsquigarrow P \in \mathcal{P}$ and $t_n \downarrow 0$, by uniform continuity of h with respect to θ from Lemma 2.A.5, and by convergence of the optimal solutions to a unique optimal solution (by Assumptions 2.3.2 and 2.3.3). This latter fact follows from continuity of the optimal solutions and optimal Lagrange multipliers, which follows from Lemma 2.A.2.

For the reverse inequality, recall the "unperturbed" and "perturbed" problems given in (2.53) and (2.54) respectively. By Lemma 2.A.1 the set of optimal solutions to program (2.54) is nonempty for all $n \geq N$. Thus, let $\theta^*_{\ell b}(P_n + t_n h_n)$ be a sequence of optimal solutions to program (2.54). Furthermore, by Lemma 2.A.1, the set of optimal Lagrange multipliers to program (2.53) is nonempty for all $n \geq N$. Now note for any $\lambda^*_{\ell b}(P_n)$ we have:

$$\Psi_I^{\ell b}(P_n) \leq \mathcal{L}(\theta^*_{\ell b}(P_n + t_n h_n), \lambda^*_{\ell b}(P_n))(P_n), \tag{2.63}$$

$$\Psi_I^{\ell b}(P_n + t_n h_n) \geq \mathcal{L}(\theta^*_{\ell b}(P_n + t_n h_n), \lambda^*_{\ell b}(P_n))(P_n + t_n h_n). \tag{2.64}$$

It follows that for $n \geq N$:

$$\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n)$$

$$\geq \mathcal{L}(\theta^*_{\ell b}(P_n + t_n h_n), \lambda^*_{\ell b}(P_n))(P_n + t_n h_n) - \mathcal{L}(\theta^*_{\ell b}(P_n + t_n h_n), \lambda^*_{\ell b}(P_n))(P_n) \qquad \text{(by (2.63) and (2.64))}$$

$$= t_n h_{n,1} \psi(W, \theta^*_{\ell b}(P_n + t_n h_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P_n) t_n h_{n,j+1} m_j(W, \theta^*_{\ell b}(P_n + t_n h_n)) \qquad \text{(by (2.52))}$$

$$= t_n \left(h_1 \psi(W, \theta^*_{\ell b}(P_n + t_n h_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P_n) h_{j+1} m_j(W, \theta^*_{\ell b}(P_n + t_n h_n)) \right) + o(t_n), \quad \text{(by (2.61))}$$

where the final line follows from uniform boundedness of the Lagrangian from Lemma 2.A.3(ii). Thus,

$$\liminf_{n\to\infty} \frac{\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n)}{t_n}$$

$$\geq \liminf_{n\to\infty} \left[h_1 \psi(W, \theta^*_{\ell b}(P_n + t_n h_n)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P_n) h_{j+1} m_j(W, \theta^*_{\ell b}(P_n + t_n h_n)) \right]$$

$$= h_1 \psi(W, \theta^*_{\ell b}(P)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P) h_{j+1} m_j(W, \theta^*_{\ell b}(P)). \tag{2.65}$$

The last line follows by convergence of $P_n \rightsquigarrow P \in \mathcal{P}$ and $t_n \downarrow 0$, by uniform continuity of h with respect to θ from Lemma 2.A.5, and by convergence of the optimal solutions to a unique optimal solution (by Assumptions 2.3.2 and 2.3.3). This latter fact follows from continuity of the optimal solutions and optimal Lagrange multipliers, which follows from Lemma 2.A.2.

Finally, combining inequalities we obtain:

$$\lim_{n\to\infty} \frac{\Psi_I^{\ell b}(P_n + t_n h_n) - \Psi_I^{\ell b}(P_n)}{t_n} = h_1 \psi(W, \theta^*_{\ell b}(P)) + \sum_{j=1}^{k} \lambda^*_{\ell b,j}(P) h_{j+1} m_j(W, \theta^*_{\ell b}(P)). \tag{2.66}$$

This completes the proof.

■

Proof of Lemma 2.3.1. Let $\mathbb{G}_{n,P_n} = \sqrt{n}(\mathbb{P}_n - P_n)$. By Lemma D.1(2) in Bugni et al. (2015) we have that, under Assumptions 2.3.1 and 2.3.5, $\mathcal{F}$ is Donsker and pre-Gaussian, both uniformly over $\mathcal{P}$. By Theorem 2.8.7 in Van Der Vaart and Wellner (1996), we have that Assumption 2.3.1 and 2.3.5 imply that $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$, which is a tight Gaussian process with sample paths that are almost all uniformly continuous. Let $\tilde{\mathbb{G}}_P$ be a version of $\mathbb{G}_P$ with all sample paths uniformly continuous. Let $\mathbb{D} = \ell^\infty(\mathcal{F})$, $\mathbb{D}_0 = \mathcal{T}_P(\mathcal{F})$, $\mathbb{E} = \mathbb{R}$, and define:

$$\mathbb{D}_n = \{h : P_n + n^{-1/2} h \in \mathcal{P}_+\}.$$

Then $\mathbb{D}_n \subset \mathbb{D}$ and $\mathbb{D}_0 \subset \mathbb{D}$. Now consider the maps $g_n : \mathbb{D}_n \to \mathbb{E}$ and $g : \mathbb{D}_0 \to \mathbb{E}$ defined as:

$$g_n(h_n) := \sqrt{n}\left\{\Psi_I^{\ell b}\left(P_n + n^{-1/2}h_n\right) - \Psi_I^{\ell b}\left(P_n\right)\right\}, \qquad h_n \in \mathbb{D}_n, \tag{2.67}$$

$$g(h) := (\Psi_I^{\ell b})'_P(h), \qquad h \in \mathbb{D}. \tag{2.68}$$

By Theorem 2.3.1, if $h_n \to h$ with $h_n \in \mathbb{D}_n$ for every n and $h \in \mathbb{D}_0$, then $g_n(h_n) \to g(h)$, where $g : \mathbb{D}_0 \to \mathbb{D}$. Now note that $\mathbb{G}_{n,P_n} \in \mathbb{D}_n$. Using the fact that $\tilde{\mathbb{G}}_P$ is a tight (and thus separable) Borel element with values in $\mathbb{D}_0$, combined with the extended continuous mapping theorem (Theorem 1.11.1 in Van Der Vaart and Wellner (1996)), we conclude that:

$$\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \rightsquigarrow (\Psi_I^{\ell b})'_P(\tilde{\mathbb{G}}_P),$$

as desired. An identical proof can be completed for the upper bound. Thus, this completes the proof. ■

Proof of Lemma 2.3.2. Let $\tilde{\mathbb{G}}_P$ be a version of $\mathbb{G}_P$ with all sample paths uniformly continuous. Let $\mathbb{D} = \ell^\infty(\mathcal{F})$, $\mathbb{D}_0 = \mathcal{T}_P(\mathcal{F})$, $\mathbb{E} = \mathbb{R}$, and define:

$$\mathbb{D}_n = \{h \in \mathbb{D} : \mathbb{P}_n + n^{-1/2}h \in \mathcal{P}_+\},$$

Then $\mathbb{D}_n \subset \mathbb{D}$ and $\mathbb{D}_0 \subset \mathbb{D}$. Now consider the maps $g_n : \mathbb{D}_n \to \mathbb{E}$ and $g : \mathbb{D}_0 \to \mathbb{E}$ defined as:

$$g_n(h_n) := \sqrt{n}\left(\Psi_I^{\ell b}\left(\mathbb{P}_n + n^{-1/2}h_n\right) - \Psi_I^{\ell b}\left(\mathbb{P}_n\right)\right), \qquad h_n \in \mathbb{D}_n, \tag{2.69}$$

$$g(h) := (\Psi_I^{\ell b})'_P(h), \qquad h \in \mathbb{D}. \tag{2.70}$$

By Theorem 2.3.1, if $h_n \to h$ with $h_n \in \mathbb{D}_n$ for every n and $h \in \mathbb{D}_0$, then $g_n(h_n) \to g(h)$, where $g : \mathbb{D}_0 \to \mathbb{D}$. Now note that $\mathbb{G}_n^b \in \mathbb{D}_n$, and by assumption $\mathbb{G}_n^b | \{W_i\}_{i=1}^n \rightsquigarrow \tilde{\mathbb{G}}_P$ uniformly over $\mathcal{P}$. Using the fact that $\tilde{\mathbb{G}}_P$ is a tight (and thus separable) Borel element with values in $\mathbb{D}_0$, combined with the extended continuous mapping theorem (Theorem 1.11.1 in Van Der Vaart and Wellner (1996)), we conclude that:

$$\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n))|\{W_i\}_{i=1}^n \rightsquigarrow (\Psi_I^{\ell b})'_P(\tilde{\mathbb{G}}_P),$$

as desired. An identical proof can be completed for the upper bound. Thus, this completes the proof.

■

Proof of Theorem 2.3.2. By definition there exists a sequence $(\psi_n, P_n) \in \{(\psi, P) : \psi \in \Psi_I(P), \ P \in \mathcal{P}\}$ satisfying:

$$\liminf_{n\to\infty} \inf_{\{(\psi,P):\ \psi\in\Psi_I(P),\ P\in\mathcal{P}\}} P\left(\psi \in C_n^\psi(1-\alpha)\right) = \liminf_{n\to\infty} P_n\left(\psi_n \in C_n^\psi(1-\alpha)\right),$$

where $\{\psi_n\}$ is a sequence with $\psi_n \in [\Psi_I^{\ell b}(P_n), \Psi_I^{ub}(P_n)]$ for each n. For such a sequence, there exists a convergent subsequence indexed by n' such that:

$$\liminf_{n\to\infty} P_n\left(\psi_n \in C_n^\psi(1-\alpha)\right) = \lim_{n'\to\infty} P_{n'}\left(\psi_{n'} \in C_{n'}^\psi(1-\alpha)\right).$$

Under our assumptions $\mathcal{P}$ is closed and uniformly tight; thus, by extracting a further subsequence if necessary, we can assume that $P_{n'} \rightsquigarrow P$ for some $P \in \mathcal{P}$. For the remainder of the proof we will argue along

this subsequence, and abusing notation we will refer to this subsequence by n rather than n'. Since by construction we have $\psi_n \in [\Psi_I^{\ell b}(P_n), \Psi_I^{ub}(P_n)]$ for each n, it suffices to establish that:

$$\lim_{n\to\infty} P_n\left(\Psi_I^{\ell b}(P_n) \in C_n^{\psi}(1-\alpha)\right) \geq 1-\alpha, \tag{2.71}$$

$$\lim_{n\to\infty} P_n\left(\Psi_I^{ub}(P_n) \in C_n^{\psi}(1-\alpha)\right) \geq 1-\alpha. \tag{2.72}$$

We can focus on (2.71) since (2.72) can be treated analogously. We have:

$$\begin{aligned}
&P_n\left(\Psi_I^{\ell b}(P_n) \in C_n^{\psi}(1-\alpha)\right) \\
&= P_n\left(\Psi_I^{\ell b}(\mathbb{P}_n) - \hat{\Psi}_\alpha^{\ell b}/\sqrt{n} \leq \Psi_I^{\ell b}(P_n) \leq \Psi_I^{ub}(\mathbb{P}_n) + \hat{\Psi}_\alpha^{ub}/\sqrt{n}\right) \\
&= P_n\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n))\right) \\
&= P_n\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P_n)) + \sqrt{n}\Delta(P_n)\right).
\end{aligned} \tag{2.73}$$

Decomposing this probability we have:

$$\begin{aligned}
&P_n\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P_n)) + \sqrt{n}\Delta(P_n)\right) \\
&= P_n^b\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)) + \sqrt{n}\Delta(\mathbb{P}_n)\right) \\
&+ \left[P_n\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P_n)) + \sqrt{n}\Delta(P_n)\right)\right. \\
&\quad \left. - P_n\left((\Psi_I^{\ell b})'_P(\mathbb{G}_P) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq (\Psi_I^{ub})'_P(\mathbb{G}_P) + \sqrt{n}\Delta(P_n)\right)\right]
\end{aligned} \tag{2.74}$$

$$\begin{aligned}
&+ \left[P_n\left((\Psi_I^{\ell b})'_P(\mathbb{G}_P) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq (\Psi_I^{ub})'_P(\mathbb{G}_P) + \sqrt{n}\Delta(P_n)\right)\right. \\
&\quad \left. - P_n\left((\Psi_I^{\ell b})'_P(\mathbb{G}_P) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq (\Psi_I^{ub})'_P(\mathbb{G}_P) + \sqrt{n}\Delta(\mathbb{P}_n)\right)\right]
\end{aligned} \tag{2.75}$$

$$\begin{aligned}
&+ \left[P_n\left((\Psi_I^{\ell b})'_P(\mathbb{G}_P) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq (\Psi_I^{ub})_P(\mathbb{G}_P) + \sqrt{n}\Delta(\mathbb{P}_n)\right)\right. \\
&\quad \left. - P_n^b\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)) + \sqrt{n}\Delta(\mathbb{P}_n)\right)\right].
\end{aligned} \tag{2.76}$$

Note by construction we will have for all n:

$$P_n^b\left(\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n^b) - \Psi_I^{\ell b}(\mathbb{P}_n)) \leq \hat{\Psi}_\alpha^{\ell b},\ -\hat{\Psi}_\alpha^{ub} \leq \sqrt{n}(\Psi_I^{ub}(\mathbb{P}_n^b) - \Psi_I^{ub}(\mathbb{P}_n)) + \sqrt{n}\Delta(\mathbb{P}_n)\right) \geq 1-\alpha,$$

so that it suffices to show that the terms (2.74), (2.75) and (2.76) converge to non-negative values. First

consider (2.74). By Lemma 2.3.1 we have that:

$$\sqrt{n}(\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P_n)) \rightsquigarrow (\Psi_I^{\ell b})'_P(\mathbb{G}_P). \tag{2.77}$$

Assumptions 2.3.1 and 2.3.5 ensure the objective function (when it is a non-trivial function of $w \in \mathcal{W}$) and moment functions are uniformly Donsker over $\mathcal{P}$. Thus, when combined with uniform boundedness of the Lagrange multipliers from Lemma 2.A.3, this ensures continuity of the distribution of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ at its α quantile and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$ at its $1-\alpha$ quantile. Thus, convergence of (2.74) to zero follows from (2.77), Theorem 1.3.4(vi) in Van Der Vaart and Wellner (1996), and continuity of the distributions of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$.

Next, note from Lemma 2.A.4 that $\Psi_I(\mathbb{P}_n)$ is Hausdorff consistent for $\Psi_I(P)$ over $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$, which implies consistency of $\Delta(\mathbb{P}_n)$ for $\Delta(P)$. Also note that Assumptions 2.3.2 and 2.3.3 imply that $\Delta(P) > 0$ for all $P \in \mathcal{P}$, so that $\sqrt{n}\Delta(P_n) \to \infty$. However, $\Delta(\mathbb{P}_n) = \Delta(P) + o_{P_n}(1)$ by Lemma 2.A.4, so that (2.75) converges to zero, as desired.

Finally, (2.76) converges to zero w.p.a. 1, which follows from bootstrap consistency over the sequence $\{P_n \in \mathcal{P}\}_{n=1}^{\infty}$ from Lemma 2.3.2, and again from continuity of the distributions of $(\Psi_I^{\ell b})'_P(\mathbb{G}_P)$ and $(\Psi_I^{ub})'_P(\mathbb{G}_P)$ described above.

■

2.A.2 Proofs of Additional Results

Lemma 2.A.1. *Under Assumptions 2.3.1 - 2.3.3, $\Theta_I(P)$, $\Theta_{\ell b}(P)$, $\Theta_{ub}(P)$, $\Lambda_{\ell b}(P)$ and $\Lambda_{ub}(P)$ are nonempty for every $P \in \mathcal{P}$. Furthermore, if $\{P_n\}_{n=1}^{\infty}$ is any sequence weakly converging to $P \in \mathcal{P}$, then there exists an N such that $\Theta_I(P_n)$, $\Theta_{\ell b}(P_n)$, $\Theta_{ub}(P_n)$, $\Lambda_{\ell b}(P_n)$ and $\Lambda_{ub}(P_n)$ are nonempty for all $n \geq N$.*

Proof. Nonemptiness of $\Theta_I(P)$, $\Theta_{\ell b}(P)$, $\Theta_{ub}(P)$ follows from Assumption 2.3.1. Nonemptiness of $\Lambda_{\ell b}(P)$ and $\Lambda_{ub}(P)$ follows from 2.3.3(ii) and Wachsmuth (2013) Theorems 1 and 2.

The second claim can be established from 2.3.3(ii) and Wachsmuth (2013) Theorems 1 and 2 if we can show there exists an N such that $\Theta_I(P_n)$ is nonempty for all $n \geq N$. This follows immediately from Assumption 2.3.3(i) and the definition of convergence of probability measures used in this chapter. ■

Lemma 2.A.2. *Under Assumptions 2.3.1, 2.3.2, and 2.3.3, we have:*

(a) $\theta_{\ell b}^*(P)$ and $\theta_{ub}^*(P)$ are continuous at any $P \in \mathcal{P}$.

(b) $\lambda_{\ell b}^*(P)$ and $\lambda_{ub}^*(P)$ are continuous at any $P \in \mathcal{P}$.

(c) $\Psi_I^{\ell b}(P)$ and $\Psi_I^{ub}(P)$ are continuous at any $P \in \mathcal{P}$.

Proof. Let $||x-y||_{\overline{\mathbb{R}}_+^k} = ||\arctan(x) - \arctan(y)||$, where $||\cdot||$ is the euclidean norm. Note that $(\Theta, ||\cdot||)$, $(\overline{\mathbb{R}}_+^k, ||\cdot||_{\overline{\mathbb{R}}_+^k})$ and $(\mathcal{P}_+, d_{BL})$ are all metric spaces. Focus first on the lower bound program in (2.27). Take any $P \in \mathcal{P}$. Define:

$$\Theta_I(P_n) := \{\theta \in \Theta : P_n m_j(W,\theta) = 0,\ j = 1,\ldots,r_1,\ P_n m_j(W,\theta) \leq 0,\ j = r_1+1,\ldots,r_1+r_2\}.$$

By Lemma 2.A.1, for any sequence $P_n \rightsquigarrow P \in \mathcal{P}$ (possibly with $P_n \in \mathcal{P}_+$) we have that there exists an N such that $\Theta_I(P_n)$ is nonempty for all $n \geq N$. By Assumption 2.3.1(i), $\Theta_I(\cdot)$ is also a compact-valued correspondence for all $n \geq N$. Recall the Lagrangian for problem (2.27):

$$\mathcal{L}(\theta,\lambda)(P) := P\psi(W,\theta) + \sum_{j=1}^{k} \lambda_j P m_j(W,\theta).$$

By Assumption 2.3.2, $\mathcal{L}(\theta, \lambda)(P)$ is continuous in (θ, λ, P). Define:

$$\Theta^*(\lambda, P) := \arg\min\{\mathcal{L}(\theta, \lambda, P) : \ \theta \in \Theta_I(P)\},$$

$$\mathcal{L}^*_\theta(\lambda, P) := \min\{\mathcal{L}(\theta, \lambda, P) : \ \theta \in \Theta_I(P)\}.$$

Note that $\Theta^*(\lambda, P) \neq \emptyset$ and $\mathcal{L}^*_\theta(\lambda, P) > -\infty$ by Lemma 2.A.1. By the Theorem of the Maximum (Ok (2007), p. 306) we have that $\Theta^*(\lambda, P)$ is compact-valued, upper-hemicontinuous, and closed, and the profiled-Lagrangian $\mathcal{L}^*_\theta(\lambda, P)$ is continuous in (λ, P). Now define:

$$\Lambda^*_\theta(P) := \arg\max\{\mathcal{L}^*_\theta(\lambda, P) : \ \lambda \in \overline{\mathbb{R}}^k_+\},$$

$$\mathcal{L}^*_{\theta,\lambda}(P) := \max\{\mathcal{L}^*_\theta(\lambda, P) : \ \lambda \in \overline{\mathbb{R}}^k_+\}.$$

Note that $\Lambda^*_\theta(P) \neq \emptyset$ and $\mathcal{L}^*_{\theta,\lambda}(P) < \infty$ by Lemma 2.A.1. Applying the Theorem of the Maximum again, we have that $\Lambda^*_\theta(P)$ is compact-valued, upper-hemicontinuous, and closed, and the profiled-Lagrangian $\mathcal{L}^*_{\theta,\lambda}(P)$ is continuous in P. Similarly, define:

$$\Lambda^*(\theta, P) := \arg\max\{\mathcal{L}(\theta, \lambda, P) : \ \lambda \in \overline{\mathbb{R}}^k_+\},$$

$$\mathcal{L}^*_\lambda(\theta, P) := \max\{\mathcal{L}(\theta, \lambda, P) : \ \lambda \in \overline{\mathbb{R}}^k_+\},$$

$$\Theta^*_\lambda(P) := \arg\min\{\mathcal{L}^*_\lambda(\theta, P) : \ \theta \in \Theta_I(P)\},$$

$$\mathcal{L}^*_{\lambda,\theta}(P) := \min\{\mathcal{L}^*_\lambda(\theta, P) : \ \theta \in \Theta_I(P)\}.$$

I.e. reverse the order of profiling of the Lagrangian with respect to λ and θ. Note this can be done by strong duality (Fact 2.A.1) without affecting the optimal solution sets. Applying Lemma 2.A.1 as above, we conclude that $\Lambda^*(\theta, P) \neq \emptyset$, $\mathcal{L}^*_\lambda(\theta, P) > -\infty$, $\Theta^*_\lambda(P) \neq \emptyset$, and $\mathcal{L}^*_{\lambda,\theta}(P) < \infty$. Applying the Theorem of the Maximum sequentially as above, we conclude that $\Theta^*_\lambda(P)$ is compact-valued, upper-hemicontinuous, and closed, and the profiled-Lagrangian $\mathcal{L}^*_{\lambda,\theta}(P)$ is continuous in P. Finally, by strong duality (Fact 2.A.1) we conclude $\Psi^{\ell b}_I(P) = \mathcal{L}^*_{\theta,\lambda}(P) = \mathcal{L}^*_{\lambda,\theta}(P)$, $\Lambda_{\ell b}(P) = \Lambda^*_\theta(P)$, and $\Theta_{\ell b}(P) = \Theta^*_\lambda(P)$. By Assumption 2.3.3, all of these sets are singletons. Repeating the excercise for the upper bound program, the proof is complete. ■

Lemma 2.A.3. *Under Assumptions 2.3.1, 2.3.2, 2.3.3 and 2.3.4,*

(i) There exists constants $L_{\ell b}, L_{ub} < \infty$ such that:

$$\sup_{P \in \mathcal{P}^\varepsilon} ||\lambda^*_{\ell b}(P)|| \leq L_{\ell b}, \tag{2.78}$$

$$\sup_{P \in \mathcal{P}^\varepsilon} ||\lambda^*_{ub}(P)|| \leq L_{ub}. \tag{2.79}$$

I.e. the Lagrange multipliers are uniformly bounded over $\mathcal{P}$ in both the lower bound and upper bound programs.

(ii) There exist constants $C_{\ell b}, C_{ub} < \infty$ such that:

$$\sup_{P\in\mathcal{P}^\varepsilon}\left|\psi(W,\theta^*_{\ell b}(P))+\sum_{j=1}^{k}\lambda^*_{\ell b,j}(P)m_j(W,\theta^*_{\ell b}(P))\right|\leq C_{\ell b}, \tag{2.80}$$

$$\sup_{P\in\mathcal{P}^\varepsilon}\left|\psi(W,\theta^*_{ub}(P))+\sum_{j=1}^{k}\lambda^*_{ub,j}(P)m_j(W,\theta^*_{ub}(P))\right|\leq C_{ub}. \tag{2.81}$$

I.e. the Lagrangian is uniformly bounded in both the lower bound and upper bound programs.

Proof. Part (i): We will focus on (2.78) since (2.79) follows analogously. By Assumption 2.3.2 and 2.3.3, we have the KKT conditions:

$$\nabla_\theta P\boldsymbol{m}(W,\theta^*_{\ell b}(P))^T\boldsymbol{\lambda}^*_{\ell b}(P)=-\nabla_\theta P\psi(W,\theta^*_{\ell b}(P))^T,$$

where $\nabla_\theta P\boldsymbol{m}(W,\theta^*_{\ell b}(P))$ is the $(r_1+r_2)\times d_\theta$ Jacobian matrix for the moment conditions. Let B denote the index set for the active constraints. Now let $\boldsymbol{\lambda}_B(P)$ denote the subvector of $\boldsymbol{\lambda}^*_{\ell b}(P)$ corresponding to the active constraints. Then clearly:

$$\nabla_\theta P\boldsymbol{m}(W,\theta^*_{\ell b}(P))^T\boldsymbol{\lambda}^*_{\ell b}(P)=\boldsymbol{G}(\theta^*_{\ell b}(P),P)^T\boldsymbol{\lambda}_B(P)=-\nabla_\theta P\psi(W,\theta^*_{\ell b}(P))^T.$$

Pre-multiplying by $\boldsymbol{G}(\theta^*_{\ell b}(P),P)$ and inverting (made possible by Assumption 2.3.3(ii)) we obtain:

$$\boldsymbol{\lambda}_B(P)=-\left[\boldsymbol{G}(\theta^*_{\ell b}(P),P)\boldsymbol{G}(\theta^*_{\ell b}(P),P)^T\right]^{-1}\boldsymbol{G}(\theta^*_{\ell b}(P),P)\nabla_\theta P\psi(W,\theta^*_{\ell b}(P))^T.$$

Denote:

$$\boldsymbol{A}_1(P):=-\left[\boldsymbol{G}(\theta^*_{\ell b}(P),P)\boldsymbol{G}(\theta^*_{\ell b}(P),P)^T\right]^{-1}, \tag{2.82}$$

$$\boldsymbol{A}_2(P):=\boldsymbol{G}(\theta^*_{\ell b}(P),P)\nabla_\theta P\psi(W,\theta^*_{\ell b}(P))^T. \tag{2.83}$$

Now note:

$$\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_1(P)||_2\leq\frac{1}{\sqrt{\kappa}},\qquad\text{(by Assumption 2.3.3)},$$

$$\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_2(P)||\leq\sqrt{\kappa}\cdot L_{\ell b},\qquad\text{(by Assumption 2.3.4)},$$

where $||\cdot||_2$ denotes the 2−matrix norm and $L_{\ell b}<\infty$ is some constant. Then:

$$\begin{aligned}\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{\lambda}_B(P)||&=\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_1(P)\boldsymbol{A}_2(P)||\\&=\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_1(P)\boldsymbol{A}_2(P)||_F\\&\leq\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_1(P)||_2||\boldsymbol{A}_2(P)||_F\\&\leq\left(\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_1(P)||_2\right)\left(\sup_{P\in\mathcal{P}^\varepsilon}||\boldsymbol{A}_2(P)||_F\right)\\&\leq L_{\ell b},\end{aligned}$$

where $||\cdot||_F$ denotes the Frobenius norm. After some transformation, this upper bound is also sufficient for

the arctan norm, and completes the proof of the first part.

Part (ii): We will focus on (2.80) since (2.81) follows analogously. By Assumption 2.3.1(v) there exists a function $F(w)$ such that $\sup_{\theta\in\Theta}||f(w,\theta)|| \leq ||F(w)||$ for every $w \in \mathcal{W}$, and such that $F(w)$ is uniformly bounded. Let $C_F < \infty$ be a positive constant satisfying $||F(w)|| \leq C_F$ for all $w \in \mathcal{W}$. Then:

$$\begin{aligned}\sup_{P\in\mathcal{P}^\varepsilon}\left|\psi(W,\theta^*_{\ell b}(P)) + \sum_{j=1}^{k}\lambda^*_{\ell b,j}(P)m_j(W,\theta^*_{\ell b}(P))\right| &\leq \sup_{P\in\mathcal{P}^\varepsilon}||F(w)||\cdot||\lambda^*_{\ell b}(P)|| \\ &\leq C_F L_{\ell b},\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz, and the last inequality follows from part (i). Thus, taking $C_{\ell b} = C_F L_{\ell b}$ the proof is complete. ■

Lemma 2.A.4. *Under Assumptions 2.3.1-2.3.5, we have that,*

(i) $d_H(\Theta_I(\mathbb{P}_n),\Theta_I(P)) = O_{\mathcal{P}}(n^{-1/2})$.

(ii) $d_H(\Psi_I(\mathbb{P}_n),\Psi_I(P)) = o_{\mathcal{P}}(1)$.

(iii) For any $\varepsilon > 0$,

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P_P^*\left(||\theta^*_{\ell b}(\mathbb{P}_n) - \theta^*_{\ell b}(P)|| > \varepsilon\right) = 0, \tag{2.84}$$

and the analogous result for $\theta^*_{ub}(\cdot)$.

(iv) For any $\varepsilon > 0$,

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P_P^*\left(||\lambda^*_{\ell b}(\mathbb{P}_n) - \lambda^*_{\ell b}(P)||_{\overline{\mathbb{R}}^k} > \varepsilon\right) = 0. \tag{2.85}$$

and the analogous result for $\lambda^*_{ub}(\cdot)$.

Proof of Lemma 2.A.4. Part (i): We follows closely the proof of Theorem 4.3(II) in Kaido et al. (2019a). Define the set:

$$\Theta_I^\gamma(P) := \left\{\theta\in\Theta : \max_{r_1+1\leq j\leq r_1+r_2} Pm_j(W,\theta) \leq \gamma,\ \ Pm_j(W,\theta) = 0,\ \ j = 1,\ldots,r_1\right\},$$

for $\gamma\in\mathbb{R}$. First note that by Lemma D.1 in Bugni et al. (2015) Assumption 2.3.5 implies that $\mathcal{F}$ is uniformly Donsker. In particular, we have that $||\mathbb{G}_{n,P}||_{\mathcal{F}} = O_{\mathcal{P}}(1)$. This implies:

$$\begin{aligned}\sup_{\theta\in\Theta_I^{-\varepsilon_n}(P)}\sqrt{n}\max_j|\mathbb{P}_n m_j(W,\theta)|_+ &\leq \sup_{\theta\in\Theta_I^{-\varepsilon_n}(P)}\sum_j\sqrt{n}|\mathbb{P}_n m_j(W,\theta)|_+ \\ &= \sup_{\theta\in\Theta_I^{-\varepsilon_n}(P)}\sum_j|\mathbb{G}_{n,P}m_j(W,\theta) + \sqrt{n}Pm_j(W,\theta)|_+ \\ &\leq r_1|O_{\mathcal{P}}(1)| + r_2|O_{\mathcal{P}}(1) - \sqrt{n}\varepsilon_n|_+,\end{aligned}$$

from which we conclude that $\Theta_I^{-\varepsilon_n}(P)\subseteq\Theta_I(\mathbb{P}_n)$ w.p.a. 1 for $\varepsilon_n = O_{\mathcal{P}}(n^{-1/2})$. Furthermore, by Assumption 2.3.5(iv) we can choose $\delta(\varepsilon_n) > 0$ such that:

$$\begin{aligned}&\inf_{\theta\in\Theta\setminus\Theta_I^{\varepsilon_n}(P)}\sqrt{n}\max_j|\mathbb{P}_n m_j(W,\theta)|_+ \\ &= \inf_{\theta\in\Theta\setminus\Theta_I^{\varepsilon_n}(P)}\max_j|\mathbb{G}_{n,P}m_j(W,\theta) + \sqrt{n}Pm_j(W,\theta)|_+\end{aligned}$$

$$
\begin{aligned}
&\geq \inf_{\theta \in \Theta \setminus \Theta_I^{\varepsilon_n}(P)} \frac{1}{J} \sum_j |\mathbb{G}_{n,P} m_j(W,\theta) + \sqrt{n} P m_j(W,\theta)|_+ \\
&\geq \inf_{\theta \in \Theta \setminus \Theta_I^{\varepsilon_n}(P)} \frac{1}{J} \left[(J-1) \cdot 0 + |O_{\mathcal{P}}(1) + \sqrt{n} C \min\{\delta(\varepsilon_n), d(\theta, \Theta_I(P))\}|_+ \right] \\
&= \inf_{\theta \in \Theta \setminus \Theta_I^{\varepsilon_n}(P)} \frac{1}{J} |O_{\mathcal{P}}(1) + \sqrt{n} C \min\{\delta(\varepsilon_n), d(\theta, \Theta_I(P))\}|_+,
\end{aligned}
$$

from which we conclude that $\Theta_I(\mathbb{P}_n) \cap (\Theta \setminus \Theta_I^{\varepsilon_n}(P)) = \emptyset$ w.p.a. 1 for $\varepsilon_n = O_{\mathcal{P}}(n^{-1/2})$ (from the first line). Note that this concludes the proof of part (i).

Part (ii): It suffices to show consistency of the upper and lower bounds; i.e. that $|\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P)| = o_{\mathcal{P}}(1)$ and that $|\Psi_I^{ub}(\mathbb{P}_n) - \Psi_I^{ub}(P)| = o_{\mathcal{P}}(1)$. We will focus on the lower bounds, since the upper bound proof is symmetric. First note that since $\psi(W,\theta)$ is continuous with respect to θ by Assumption 2.3.2, and that Θ is compact by Assumption 2.3.1(i), we have that $\psi(W,\theta)$ is uniformly continuous (w.r.t. θ) on Θ. Thus, for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that $|\mathbb{P}_n \psi(W,\theta) - \mathbb{P}_n \psi(W,\theta')| < \varepsilon$ whenever $||\theta - \theta'|| < \delta(\varepsilon)$.

Now note that:

$$
\begin{aligned}
|\Psi_I^{\ell b}(\mathbb{P}_n) - \Psi_I^{\ell b}(P)| &= \left| \min_{\theta \in \Theta_I(\mathbb{P}_n)} \mathbb{P}_n \psi(W,\theta) - \min_{\theta \in \Theta_I(P)} P\psi(W,\theta) \right| \\
&\leq \left| \min_{\theta \in \Theta_I(\mathbb{P}_n)} \mathbb{P}_n \psi(W,\theta) - \min_{\theta \in \Theta_I(P)} \mathbb{P}_n \psi(W,\theta) \right| + \left| \min_{\theta \in \Theta_I(P)} \mathbb{P}_n \psi(W,\theta) - \min_{\theta \in \Theta_I(P)} P\psi(W,\theta) \right| \\
&\leq \sup_{||\theta - \theta'|| \leq d_H(\Theta_I(\mathbb{P}_n), \Theta_I(P))} |\mathbb{P}_n \psi(W,\theta) - \mathbb{P}_n \psi(W,\theta')| + \sup_{\theta \in \Theta_I(P)} |\mathbb{P}_n \psi(W,\theta) - P\psi(W,\theta)|.
\end{aligned}
$$

It suffices to show the two terms in the last line of the previous array converge to zero in probability uniformly. Note that by part (i) of this Lemma, we have $d_H(\Theta_I(\mathbb{P}_n), \Theta_I(P)) = o_{\mathcal{P}}(1)$. Thus by uniform continuity of $\mathbb{P}_n \psi(W,\theta)$:

$$
\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} P^* \left(\sup_{||\theta - \theta'|| \leq d_H(\Theta_I(\mathbb{P}_n), \Theta_I(P))} |\mathbb{P}_n \psi(W,\theta) - \mathbb{P}_n \psi(W,\theta')| > \varepsilon \right) = 0.
$$

Also, by the uniform Donsker property:

$$
\sup_{\theta \in \Theta_I(P)} |\mathbb{P}_n \psi(W,\theta) - P\psi(W,\theta)| \leq \sup_{\theta \in \Theta} |\mathbb{P}_n \psi(W,\theta) - P\psi(W,\theta)| = o_{\mathcal{P}}(1).
$$

This completes the proof.

Part (iii) + (iv): Using Lemma 2.A.3, we can restrict λ to lie in the set $\Lambda := \{\lambda : ||\lambda|| \leq \max\{L_{\ell b}, L_{ub}\}\}$. Fix any $\varepsilon, \eta > 0$. By the uniform Donsker property we have:

$$
\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \sup_{f \in \mathcal{F}} ||\mathbb{P}_n - P||_{\mathcal{F}} = 0,
$$

which implies the following inequalities hold w.p.a. 1:

$$
\begin{aligned}
&\mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(\mathbb{P}_n))(P) < \mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(\mathbb{P}_n))(\mathbb{P}_n) + \varepsilon/3, \\
&\mathcal{L}(\theta_{\ell b}^*(P), \lambda_{\ell b}^*(\mathbb{P}_n))(\mathbb{P}_n) < \mathcal{L}(\theta_{\ell b}^*(P), \lambda_{\ell b}^*(\mathbb{P}_n))(P) + \varepsilon/3, \\
&\mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(\mathbb{P}_n))(P) > \mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(\mathbb{P}_n))(\mathbb{P}_n) - \eta/3, \\
&\mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(P))(\mathbb{P}_n) > \mathcal{L}(\theta_{\ell b}^*(\mathbb{P}_n), \lambda_{\ell b}^*(P))(P) - \eta/3.
\end{aligned}
$$

Furthermore, by optimality of $\theta^*_{\ell b}(\mathbb{P}_n)$ and $\lambda^*_{\ell b}(\mathbb{P}_n)$ we have:

$$\mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(\mathbb{P}_n))(\mathbb{P}_n) < \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(\mathbb{P}_n))(\mathbb{P}_n) + \varepsilon/3,$$
$$\mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(\mathbb{P}_n))(\mathbb{P}_n) > \mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(P))(\mathbb{P}_n) - \eta/3.$$

Combining these inequalities we obtain w.p.a. 1:

$$\mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(\mathbb{P}_n))(P) < \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(\mathbb{P}_n))(P) + \varepsilon \leq \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(P))(P) + \varepsilon,$$

$$\mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(\mathbb{P}_n))(P) > \mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(P))(P) - \eta \geq \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(P))(P) - \eta.$$

Now let B_θ and B_λ be any open balls around $\theta^*_{\ell b}(P)$ and $\lambda^*_{\ell b}(P)$, respectively, and set:

$$\varepsilon = \inf_{\Theta \cap B^c_\theta} \mathcal{L}(\theta, \lambda^*_{\ell b}(P))(P) - \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(P))(P),$$
$$\eta = \mathcal{L}(\theta^*_{\ell b}(P), \lambda^*_{\ell b}(P))(P) - \sup_{\Lambda \cap B^c_\lambda} \mathcal{L}(\theta^*_{\ell b}(P), \lambda)(P).$$

Note by Assumption 2.3.3, we have that the optimal solutions and Lagrange multipliers are unique, so that $\varepsilon, \eta > 0$. Combining with the results above we conclude that w.p.a. 1:

$$\sup_{\Lambda \cap B^c_\lambda} \mathcal{L}(\theta^*_{\ell b}(P), \lambda)(P) < \mathcal{L}(\theta^*_{\ell b}(\mathbb{P}_n), \lambda^*_{\ell b}(\mathbb{P}_n))(P) < \inf_{\Theta \cap B^c_\theta} \mathcal{L}(\theta, \lambda^*_{\ell b}(P))(P).$$

Furthermore at least one of the inequalities in the previous display is violated if either $\theta^*_{\ell b}(\mathbb{P}_n) \notin B_\theta$ or $\lambda^*_{\ell b}(\mathbb{P}_n) \notin B_\lambda$, which concludes the proof. ■

Lemma 2.A.5. *Under Assumptions 2.3.1 and 2.3.5:*

(i) For every $\varepsilon > 0$ there exists a $\delta > 0$ such that $||\theta - \theta'|| < \delta$ implies $\rho_P(\theta, \theta') < \varepsilon$ for all $P \in \mathcal{P}$.

(ii) Any function $h \in \ell^\infty(\mathcal{F})$ uniformly continuous in the sup-norm with respect to ρ_P is uniformly continuous in the sup-norm with respect to $||\cdot||$.

Proof. Part (i): Recall that under Assumption 2.3.5 the semimetric ρ_P satisfies:

$$\lim_{\delta \downarrow 0} \sup_{||(\theta_1, \theta_1') - (\theta_2, \theta_2')|| < \delta} \sup_{P \in \mathcal{P}} |\rho_P(\theta_1, \theta_1') - \rho_P(\theta_2, \theta_2')| = 0.$$

Now take $(\theta_2, \theta_2') = (\theta_1', \theta_1')$ and obtain:

$$\lim_{\delta \downarrow 0} \sup_{||\theta_1 - \theta_1'|| < \delta} \sup_{P \in \mathcal{P}} \rho_P(\theta_1, \theta_1') = 0.$$

Thus, we conclude for any $\varepsilon > 0$ there exists a $\delta > 0$ such that:

$$\sup_{||\theta_1 - \theta_1'|| < \delta} \sup_{P \in \mathcal{P}} \rho_P(\theta_1, \theta_1') < \varepsilon.$$

In other words:

$$\{\theta, \theta' \in \Theta : ||\theta - \theta'|| < \delta\} \subseteq \{\theta, \theta' \in \Theta : \rho_P(\theta, \theta') < \varepsilon\}.$$

Part (ii): By uniform continuity of h we have for any $\eta > 0$, there exists a $\varepsilon(\eta, P) > 0$ such that:

$$\sup_{\rho_P(\theta,\theta')<\varepsilon(\eta,P)} ||hf(w,\theta) - hf(w,\theta')|| < \eta.$$

However, for any such $\varepsilon(\eta, P) > 0$, by Part (i) there exists a $\delta(\varepsilon(\eta, P))) > 0$ such that:

$$\{\theta, \theta' \in \Theta : ||\theta - \theta'|| < \delta\} \subseteq \{\theta, \theta' \in \Theta : \rho_P(\theta, \theta') < \varepsilon\}.$$

We conclude that for any $\eta > 0$ there exists a $\delta(\eta, P) > 0$ such that:

$$\sup_{||\theta-\theta'||<\delta(\eta,P)} ||hf(w,\theta) - hf(w,\theta')|| \leq \sup_{\rho_P(\theta,\theta')<\varepsilon(\eta,P)} ||hf(w,\theta) - hf(w,\theta')|| < \eta,$$

which completes the proof. ■

Appendix 2.B Further Simulation Evidence

In addition to the Monte Carlo experiment performed in the main text, we now present Monte Carlo excercises for two canonical partial identification examples given by the missing data problem from Example 1 and the linear regression example with interval-valued dependent variable from Example 2. In all Monte Carlo exercises we take $B = 1000$ bootstrap samples for each experiment, and we implement each experiment 1000 times to determine the simulated coverage probability. In each DGP, we also threshold the length of the identified set; i.e. we use $\Delta_n^* = \mathbb{1}\{\Delta_n > b_n\}$, with $b_n = (\log(n))^{-1/2}$, rather than Δ_n when computing the critical values from (2.42) and (2.43). As mentioned in the main text, we find this thresholding helps to improve the coverage in finite sample in cases when the model is close to point identification, and introduces at most a conservative distortion under the assumptions in this chapter.

2.B.1 Missing Data Example

Description

Recall that in the missing data example the researcher observes a sample $\{Y_i D_i, D_i\}_{i=1}^n$. In the Monte Carlo experiments we assume that $Y_i \in \mathcal{Y} = \{1, 2, 3, 4, 5\}$ and $D_i \in \{0, 1\}$. The parameter of interest is the unconditional average outcome:

$$\psi(\theta) = \sum_{d\in\{0,1\}} \sum_{y\in\mathcal{Y}} y \cdot \theta_{yd}, \qquad \theta_{yd} = P(Y = y, D = d).$$

The constraints imposed by the observed distribution $P(YD, D)$ on the latent distribution $P(Y, D)$ are given by:

$$P(YD = y, D = 1) = P(Y = y, D = 1), \quad \forall y \in \mathcal{Y}, \tag{2.86}$$

$$P(YD = 0, D = 0) = \sum_{y\in\mathcal{Y}} P(Y = y, D = 0). \tag{2.87}$$

The identified set for ψ is given by $\Psi_I(P) = [\Psi_I^{\ell b}(P), \Psi_I^{ub}(P)]$, and can be obtained by solving the problems:

$$\Psi_I^{\ell b}(P) = \min_{\theta\in\Theta_I(P)} \psi(\theta), \qquad \Psi_I^{ub}(P) = \max_{\theta\in\Theta_I(P)} \psi(\theta), \tag{2.88}$$

where $\Theta_I(P)$ is the set of probability vectors satisfying the constraints (2.86) and (2.87). In the Monte Carlo exercise we take $n \in \{100, 250, 500, 1000\}$, and specify the DGP as:

$$P(Y = y, D = 1) = \frac{1}{5}\left(1 - \max\left\{\frac{c}{\sqrt{n}}, \delta\right\}\right), \qquad P(Y = y, D = 0) = \frac{1}{5}\max\left\{\frac{c}{\sqrt{n}}, \delta\right\}, \quad \forall y \in \mathcal{Y},$$

for $c \in \{0.1, 1, 2\}$, and for some small $\delta > 0$ (we take $\delta = 10^{-6}$).[9] Note this corresponds to a DGP where $\psi_0 = 3$, which will always be partially identified. Notice that all constraints on the identified set can be expressed $A\theta - b = 0$, where:

$$\theta = \begin{bmatrix} \theta_{10} \\ \vdots \\ \theta_{50} \\ \theta_{11} \\ \vdots \\ \theta_{51} \end{bmatrix}, \qquad A = \begin{bmatrix} \underset{(1\times5)}{\boldsymbol{\iota}'} & \underset{(1\times5)}{\mathbf{0}} \\ \underset{(5\times5)}{\mathbf{0}} & \underset{(5\times5)}{\boldsymbol{I}} \end{bmatrix}, \qquad b = \begin{bmatrix} P(D=0) \\ P(Y=1, D=1) \\ P(Y=2, D=1) \\ P(Y=2, D=1) \\ P(Y=4, D=1) \\ P(Y=5, D=1) \end{bmatrix},$$

where $\boldsymbol{\iota}$ denotes a vector of 1's, and $\boldsymbol{I}$ denotes the identity matrix.

Results

The simulation results for the missing data example are displayed in Table 2.2. As is expected under partial identification of the parameter ψ_0, the coverage probability for the true parameter is slightly above nominal. In particular, this results from the fact that often the true parameter lies interior to the identified set. Note this also occurs because of the thresholding discussed at the beginning of this section, which will introduce a slight conservative distortion under our assumptions. As the value of c increases, we see that the length of the identified set increases due to the fact that the missing data probability is increasing. However, in all of the DGPs considered our confidence sets remain informative. The linear programming formulation of this problem also ensures that the confidence set can be computed very efficiently; the approximate time to compute a confidence set was typically below 4 seconds.

2.B.2 Interval Valued Regression

Description

Recall the example of linear regression with interval-valued dependent variable. We have $Y = X\theta + \varepsilon$, where $X \in \mathbb{R}^d$ with R points of support, and values of Y are never observed, although we observe realizations of two random variables Y^* and Y_* satisfying $P(Y_* \leq Y \leq Y^*) = 1$. The objective is then to perform inference for the subvector θ_1 of θ given that researcher observes a sample $\{Y_i^*, Y_{*i}, X_i\}_{i=1}^n$. Recall the identified set is given by:

$$\Theta_I(P) := \{\theta : \mathbb{E}[Y_*|X = x_r] - x_r^T\theta \leq 0, \;\; x_r^T\theta - \mathbb{E}[Y^*|X = x_r] \leq 0, \;\; r = 1, \ldots, R\}.$$

[9] The inclusion of δ is mostly a theoretical indulgence, since it ensures that the probability of data being missing is always positive, even asymptotically. However, in our DGPs we will always have $\delta < c/\sqrt{n}$ so that practically it plays no role in our Monte Carlo study.

Setting $\psi(W,\theta) = \psi(\theta) = \theta_1$, the identified set for the functional ψ is an interval $\Psi_I(\mathbb{P}_n) = [\Psi_I^{\ell b}(\mathbb{P}_n), \Psi_I^{ub}(\mathbb{P}_n)]$ with the endpoints determined by:

$$\Psi_I^{\ell b}(\mathbb{P}_n) = \min_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta), \qquad \Psi_I^{ub}(\mathbb{P}_n) = \max_{\theta \in \Theta_I(\mathbb{P}_n)} \psi(\theta). \tag{2.89}$$

In our DGP, we set $Y = X\theta + \varepsilon$, where $X \in \{0,1\}^4$, with each component of X generated according to a Bernoulli(0.5) distribution, and where $\varepsilon \sim N(0,1)$. Note this implies that X has $R = 16$ points of support. We assume that the random variables Y_* and Y^* are generated according to:

$$Y_* = Y - \max\{c/\sqrt{n}, \delta\},$$
$$Y^* = Y + \max\{c/\sqrt{n}, \delta\},$$

for $c \in \{1, 5, 10\}$, depending on the DGP, and $\delta = 10^{-6}$. Note that the model would be point-identified if we set $c = 0$ and $\delta = 0$. Notice that all constraints on the identified set can be expressed $A\theta - b \leq 0$, where:

$$\underset{(2R\times 4)}{A} = \begin{bmatrix} -x_1^T \cdot P(X = x_1) \\ -x_2^T \cdot P(X = x_2) \\ \vdots \\ -x_R^T \cdot P(X = x_R) \\ x_1^T \cdot P(X = x_1) \\ x_2^T \cdot P(X = x_2) \\ \vdots \\ x_R^T \cdot P(X = x_R) \end{bmatrix}, \qquad \underset{(2R\times 1)}{b} = \begin{bmatrix} -\mathbb{E}[Y_* \mathbb{1}\{X = x_1\}] \\ -\mathbb{E}[Y_* \mathbb{1}\{X = x_2\}] \\ \vdots \\ -\mathbb{E}[Y_* \mathbb{1}\{X = x_R\}] \\ \mathbb{E}[Y^* \mathbb{1}\{X = x_1\}] \\ \mathbb{E}[Y^* \mathbb{1}\{X = x_2\}] \\ \vdots \\ \mathbb{E}[Y^* \mathbb{1}\{X = x_R\}] \end{bmatrix}.$$

Similar to the previous simulation exercises, we take sample sizes $n \in \{100, 250, 500, 1000\}$.

Results

The simulation results for the interval valued regression example are displayed in Table 2.3. Similar the missing data Monte Carlo, the coverage probability for the true parameter is slightly above nominal. Again, this results from the fact that often the true parameter lies interior to the identified set, as well as from the thresholding discussed at the beginning of this section. However, the coverage probability is very close to nominal (e.g. see the results for $n = 1000$ and $c = 1$). As the value of c increases, we see that the length of the identified set increases due to the fact that the interval length for the interval-valued outcome variable increases in length. However, in all of the DGPs considered our confidence set remain informative. Again, the linear programming formulation of this problem also ensures that the confidence set can be computed very efficiently; the approximate time to compute a confidence set was around 4 seconds.

Table 2.1: Table showing the results of the simulation excercise for the counterfactual policy example in Kasy (2016) with $B = 1000$ bootstrap replications for each experiment, where 1000 experiments are run to determine the coverage probability. The parameter of interest is ψ^{AB}, which is the difference in the expected treatment effect under two competing policies.

c=1												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	0.37	0.37	0.37	0.890	0.18	0.58	0.943	0.15	0.62	0.990	0.07	0.69
n=250	0.37	0.37	0.37	0.899	0.26	0.50	0.961	0.23	0.52	0.988	0.19	0.56
n=500	0.37	0.37	0.37	0.898	0.29	0.45	0.950	0.27	0.47	0.988	0.24	0.50
n=1000	0.37	0.37	0.37	0.914	0.31	0.43	0.962	0.30	0.44	0.992	0.28	0.46
c=10												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	0.37	0.07	0.56	0.995	-0.12	0.72	1.000	-0.17	0.78	1.000	-0.26	0.87
n=250	0.37	0.26	0.43	0.992	0.14	0.57	0.998	0.12	0.59	1.000	0.07	0.64
n=500	0.37	0.34	0.38	0.959	0.26	0.47	0.982	0.25	0.49	0.995	0.21	0.52
n=1000	0.37	0.37	0.37	0.923	0.31	0.43	0.963	0.30	0.44	0.993	0.28	0.46
c=20												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	0.37	-0.09	0.64	0.996	-0.26	0.76	1.000	-0.31	0.82	1.000	-0.42	0.92
n=250	0.37	-0.02	0.60	1.000	-0.14	0.71	1.000	-0.17	0.73	1.000	-0.23	0.79
n=500	0.37	0.12	0.52	1.000	0.04	0.60	1.000	0.02	0.62	1.000	-0.01	0.65
n=1000	0.37	0.26	0.43	1.000	0.20	0.50	1.000	0.19	0.51	1.000	0.17	0.53

"Avg. Estimate" and below "LB" and "UB" stands for the average value of the lower and upper bounds of the identified set, where the average is taken over all experiments. "Coverage" refers to the proportion of times the true value of ψ fell within the confidence region over all experiments.

Table 2.2: Table showing the results of the simulation excercise for the missing data example with $B = 1000$ bootstrap replications for each experiment, where 1000 experiments are run to determine the coverage probability. The parameter of interest is the unconditional average outcome Y where $Y \in \{1, 2, 3, 4, 5\}$ and where Y is observed only when $D = 1$.

c=0.1												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	3.00	2.97	3.01	0.918	2.73	3.25	0.941	2.69	3.29	0.986	2.60	3.38
n=250	3.00	2.99	3.01	0.931	2.83	3.16	0.956	2.81	3.19	0.987	2.75	3.24
n=500	3.00	2.99	3.01	0.920	2.88	3.12	0.960	2.86	3.14	0.993	2.82	3.18
n=1000	3.00	2.99	3.01	0.928	2.92	3.08	0.969	2.90	3.10	0.994	2.88	3.12
c=1												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	3.00	2.79	3.19	0.997	2.56	3.42	1.000	2.51	3.47	1.000	2.42	3.56
n=250	3.00	2.87	3.12	0.997	2.71	3.28	1.000	2.69	3.31	1.000	2.63	3.36
n=500	3.00	2.91	3.09	0.997	2.80	3.20	1.000	2.78	3.22	1.000	2.74	3.26
n=1000	3.00	2.94	3.06	0.999	2.86	3.14	0.999	2.84	3.15	1.000	2.82	3.18
c=2												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	3.00	2.59	3.39	1.000	2.40	3.58	1.000	2.34	3.64	1.000	2.24	3.74
n=250	3.00	2.74	3.25	1.000	2.62	3.37	1.000	2.58	3.41	1.000	2.52	3.47
n=500	3.00	2.82	3.18	1.000	2.71	3.28	1.000	2.69	3.30	1.000	2.65	3.35
n=1000	3.00	2.87	3.12	1.000	2.79	3.20	1.000	2.78	3.22	1.000	2.75	3.24

"Avg. Estimate" and below "LB" and "UB" stands for the average value of the lower and upper bounds of the identified set, where the average is taken over all experiments. "Coverage" refers to the proportion of times the true value of ψ fell within the confidence region over all experiments.

Table 2.3: Table showing the results of the simulation excercise for the missing data example with $B = 1000$ bootstrap replications for each experiment, where 1000 experiments are run to determine the coverage probability. The parameter of interest is the first component, θ_1, of the vector θ, where $Y = X\theta + \varepsilon$, and where Y is interval-valued with $P(Y_L \leq Y \leq Y_U) = 1$.

c=1												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	1.15	1.13	1.16	0.907	0.83	1.46	0.942	0.78	1.52	0.987	0.67	1.63
n=250	1.15	1.13	1.15	0.930	0.95	1.34	0.957	0.91	1.38	0.990	0.84	1.44
n=500	1.15	1.15	1.16	0.915	1.02	1.30	0.954	0.99	1.32	0.989	0.94	1.37
n=1000	1.15	1.14	1.15	0.911	1.05	1.25	0.955	1.03	1.27	0.991	1.00	1.30
c=5												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	1.15	1.07	1.23	0.951	0.77	1.53	0.977	0.71	1.59	0.995	0.60	1.70
n=250	1.15	1.09	1.20	0.963	0.90	1.38	0.982	0.87	1.42	0.997	0.80	1.49
n=500	1.15	1.12	1.19	0.959	0.99	1.32	0.976	0.96	1.35	0.997	0.92	1.40
n=1000	1.15	1.12	1.17	0.961	1.03	1.27	0.986	1.01	1.29	0.998	0.98	1.32
c=10												
		Avg. Estimate		$1-\alpha = 0.90$			$1-\alpha = 0.95$			$1-\alpha = 0.99$		
Sample Size	True Value	LB	UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB	Coverage	Avg. LB	Avg. UB
n=100	1.15	0.99	1.31	0.986	0.68	1.61	0.993	0.63	1.67	0.999	0.51	1.78
n=250	1.15	1.04	1.25	0.988	0.85	1.43	0.996	0.82	1.47	1.000	0.75	1.54
n=500	1.15	1.08	1.23	0.983	0.95	1.36	0.991	0.93	1.39	0.998	0.88	1.43
n=1000	1.15	1.10	1.20	0.990	1.00	1.29	0.997	0.99	1.31	0.998	0.95	1.35

"Avg. Estimate" and below "LB" and "UB" stands for the average value of the lower and upper bounds of the identified set, where the average is taken over all experiments. "Coverage" refers to the proportion of times the true value of ψ fell within the confidence region over all experiments.

Chapter 3

Policy Transforms and Learning Optimal Policies

We study the problem of choosing optimal policy rules in uncertain environments using models that may be incomplete and/or partially identified. We consider a policymaker who wishes to choose a policy to maximize a particular counterfactual quantity called a *policy transform*. We characterize *learnability* of a set of policy options by the existence of a decision rule that closely approximates the maximin optimal value of the policy transform with high probability. Sufficient conditions are provided for the existence of such a rule. However, learnability of an optimal policy is an ex-ante notion (i.e. before observing a sample), and so ex-post (i.e. after observing a sample) theoretical guarantees for certain policy rules are also provided. Our entire approach is applicable when the distribution of unobservables is not parametrically specified, although we discuss how semiparametric restrictions can be used. Finally, we show possible applications of the procedure to a simultaneous discrete choice example and a program evaluation example.

3.1 Introduction

One of the fundamental goals of econometrics is to credibly translate knowledge of underlying economic mechanisms into models that, when combined with sample data, can be used to understand the effects of counterfactual policy experiments and can help guide policy decisions. In this paper we consider the problem of making policy decisions in settings where the econometric model is partially identified and/or incomplete. The paper is motivated by the fact that credible models are needed to honestly inform policy makers on the impacts of counterfactual policies, even if credible models provide an incomplete description of the true data generating process.

Our framework is general enough to accommodate many existing structural econometric models. Our description of the environment is similar to descriptions found in Jovanovic (1989) and Chesher and Rosen (2017a), which in turn are extensions of the classical foundations for econometric modelling set forth in Koopmans et al. (1950) and Hurwicz (1950), among others. We assume the economic system under consideration manifests as a collection of random variables which can be partitioned into those that are observable—including a vector of observed endogenous variables Y and a vector of exogenous variables Z—and those that are latent or unobservable—denoted by the vector U. We refer colloquially to the variables contained in Y and Z as the "observables," and refer to the variables contained in U as the "unobservables." Unlike most of the existing literature, we do not take the distribution of U as a model primitive. This is in accordance with the perspective that the latent variable U represents the gap between what can be explained by

a theoretical model, and what must remain unexplained; that is, "errors in equations" rather than "errors in variables."[1] As we will demonstrate, such a distinction becomes especially important when performing counterfactual analyses.

The policymaker is assumed to have access to data on the observables, as well as an econometric model that describes how the observables are related to the unobservables. The model may depend on a vector of parameters $\theta \in \Theta$; here Θ is required only to be a complete and separable metric space, which permits many function spaces used in nonparametric analyses. We then let Γ represent an abstraction of the set of all possible policies under consideration by the policymaker, where $\gamma \in \Gamma$ denotes one such policy. Each hypothetical policy $\gamma \in \Gamma$ represents an intervention on the underlying existing economic system, which operates to generate the endogenous variables from the exogenous and unobserved variables. After the economic system is modified, the resulting system may now generate a new, or counterfactual distribution of the endogenous variables. Thus, by altering the underlying economic system, a policy intervention induces a change between the factual (or observed) and counterfactual (hypothetical and unobserved) distributions of the endogenous outcome variables. Latent variables are not affected by the policy, and instead serve as important links between the factual and counterfactual domains.[2] A policymaker's problem is then formulated as the problem of choosing a policy intervention that induces a counterfactual distribution of endogenous outcome variables that is favourable according to some criterion.

We denote the counterfactual endogenous outcome variables as $Y_\gamma^\star$, where the γ index is to emphasize the fact that its distribution will depend on the counterfactual policy experiment $\gamma \in \Gamma$ under consideration. Under this setup, this paper focuses on a particular class of counterfactual quantities that can be written in the following form:

$$I[\varphi](\gamma) := \int \varphi(v)\, dP_{V_\gamma}. \tag{3.1}$$

Here φ is some function, $V_\gamma := (Y_\gamma^\star, Y, Z, U)$ is a vector of all the random variables that describe the factual and counterfactual domains, P_{V_γ} denotes the distribution of V_γ, and v denotes a realization of V_γ. In particular, the operator $I[\,\cdot\,](\gamma)$ takes a function φ of the vector v of endogenous, exogenous, unobserved and counterfactual variables, and maps it to a function $I[\varphi](\gamma)$ of the policy parameter γ. For this reason, we refer to $I[\,\cdot\,](\gamma)$ as a *policy transform*. As we will show in our examples on simultaneous discrete choice and program evaluation, counterfactual objects that can be written as policy transforms include counterfactual choice probabilities, and counterfactual average effects. If a policymaker's counterfactual object of interest can be written as the policy transform of some function φ, then the resulting policy transform gives all the information the policymaker needs to compare various policies and make a policy choice.

Throughout the paper we consider a policymaker who wishes to maximize the value of the policy transform, although our analysis is equally applicable to the case when the policymaker wishes to minimize the value of the policy transform. With perfect knowledge of the distribution of the vector V_γ, the policymaker faces a trivial decision problem and can simply choose the policy γ that obtains the maximum of the policy transform $I[\varphi](\gamma)$. However, this idealized decision problem is rarely encountered in practice, and instead the policymaker may only have access to a finite sample of the observed random variables. Furthermore, even with an infinite sample the policy transform may not be identified under any credible assumptions. This will be especially true throughout our discussion, since we will not require that the distribution of the unobservables U be parametrically specified.

[1] These two explanations of the error term are documented by Morgan (1990) Chapter 6. We recommend Qin and Gilbert (2001) for a review of how attitudes towards the latent variables have evolved over time.

[2] From Pearl (2009) p. 211: "The background variables are the main carriers of information from the actual world to the hypothetical world; they serve as the "guardians of invariance" (or persistence) in the dynamic process that transforms the former into the latter."

To make progress, we model the policy decision problem as a decision under ambiguity, where we assume that the "true state of the world" belongs to a state space $\mathcal{S} \times \mathcal{P}_{Y,Z}$. Here $\mathcal{P}_{Y,Z}$ is the set of all Borel probability measures on the observable space $\mathcal{Y} \times \mathcal{Z}$. Furthermore, each $s \in \mathcal{S}$ is associated with a pair of conditional distributions $(P_{U|Y,Z}, P_{Y^*_\gamma|Y,Z,U})$. Taking a pair $(s, P_{Y,Z}) \in \mathcal{S} \times \mathcal{P}_{Y,Z}$ to be the true state, the policymaker can evaluate the policy transform in (3.1) corresponding to that state. Keeping the dependence on $P_{Y,Z}$ implicit, we denote the policy transform in state $(s, P_{Y,Z})$ as $I[\varphi](\gamma, s)$, and refer to it as the *state-dependent policy transform*. We then consider the policymaker's decision problem when she has access to a finite sample from the true distribution. Let Ψ_n denote the space of all possible $n-$samples $\{(y_i, z_i)\}_{i=1}^n$, and let $d : \Psi_n \to \Gamma$ denote a (measurable) decision rule that maps from sample realizations to policies. Before a sample $\psi \in \Psi_n$ is observed $d(\psi)$ will be a random variable, and the policymaker's problem is then translated into the problem of selecting a decision rule according to some reasonable criteria.

However, without knowledge of the true state, it is unclear how the policymaker should (in a prescriptive sense) choose among, or rank, various decision rules. One nearly self-evident requirement on any method of ranking decision rules is that the ranking should respect weak dominance; that is, if for every $P_{Y,Z} \in \mathcal{P}_{Y,Z}$ we have $I[\varphi](d'(\psi), s) \leq I[\varphi](d(\psi), s)$ a.s. for every $s \in \mathcal{S}$, then d should be preferred to d'. However, it is clear that many decisions rules will not be comparable according to this partial ordering.

To progress further, we introduce a preference relation over the space of all decision rules that is motivated from computational learning theory. In particular, fix any $\kappa \in (0, 1)$ and let $c_n(d, \kappa)$ be the smallest value satisfying:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) + c_n(d, \kappa) \geq \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) \right) \geq \kappa. \tag{3.2}$$

Then under our framework, a decision rule $d : \Psi_n \to \Gamma$ is weakly preferred to decision rule $d' : \Psi_n \to \Gamma$ at level κ and sample size n if $c_n(d, \kappa) \leq c_n(d', \kappa)$.[3] This preference relation appears to be new, and diverges (to some extent) from the existing literature on frequentist decision theory. However, its close connection to the probably approximately correct (PAC) learning framework from computational learning theory allows us to use a rich set of results from statistical learning theory and empirical process theory to study its theoretical properties. In addition, this preference relation induces a total ordering, and our first result in Section 3.2 demonstrates that, at a minimum, this preference relation respects weak dominance.

Given this preference relation, throughout the paper we will use the value $c_n(d, \kappa)$ to measure the "performance" or "quality" of a decision rule d for a given sample size n and confidence level κ. We then provide two sets of theoretical results for the policymaker's decision problem.

In the first set of results, we provide conditions on the decision problem that guarantees the existence of a decision rule d such that $c_n(d, \kappa)$ tends to zero as the sample size n becomes large. The existence of such a decision rule characterizes the notion of policy space learnability. The definition of policy space learnability appears to be new in economics, although it is adapted from the widely popular PAC learning framework from computer science proposed by Valiant (1984). Our particular analysis deals mostly with the decision theoretic generalization of the PAC learning model proposed by Haussler (1992), which is referred to as the *agnostic* PAC learning model.

We show that even in simple environments the policy space may not be learnable. In this case the policymaker's decision problem is still well-defined, but there will be theoretical limitations on how well any given policy can perform, even in large samples. We then provide sufficient conditions for learnability which are related to certain complexity measures of the class of functions in our problem; in particular, to the behaviour of covering/packing numbers and metric entropy. We define an "entropy growth condition," and

[3] See Definition 3.2.3.

we show that if certain key classes of functions in our environment satisfy this condition, then the policy space Γ is learnable. Primitive conditions for our entropy growth condition can be found in the literature on empirical processes and statistical learning. In addition to being sufficient for learnability, we also show how the condition can be used to establish rates of convergence.

However, since learnability is an *ex-ante* notion (i.e. before observing the sample), verifying learnability can be uninformative about the *ex-post* performance (i.e. after observing the sample) of a given policy rule. Thus, our second set of results provides a means for the policymaker to perform an ex-post analysis of her selected policy rule. First we study the finite sample properties of a particular decision rule, called the $\varepsilon-$maximin empirical (eME) rule, which selects a $\varepsilon-$maximizer of the worst case (over $s \in \mathcal{S}$) empirical version of $I[\varphi](\gamma, s)$. Using concentration inequalities, we provide an upper bound on the quantity $c_n(d, \kappa)$ when d is the eME rule, and we demonstrate how the upper bound is affected by various features of the decision problem.

However, the eME rule is only one particular rule, and for many reasons it may not be the policy rule selected by the policymaker. We thus turn to the problem of approximating the set of all policies $\gamma \in \Gamma$ satisfying:

$$\gamma \mapsto \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) \leq \delta, \tag{3.3}$$

with probability at least κ; note that any decision rule that selects a policy in this set will thus have $c_n(d, \kappa) \leq \delta$. We call this set of policies the "δ-level set," and we show how a procedure from the literature on excess risk bounds in statistical learning theory can be adapted to our environment to approximate the $\delta-$level set. Finally, we show that the eME decision rule selects a policy in the $\delta-$level with high probability for δ sufficiently large, providing further justification for its use. Unlike the first ex-ante analysis of learnability, all of the results comprising the ex-post analysis do not require the entropy growth condition—or any other sufficient condition for learnability—to be satisfied. Thus, they are applicable whether or not the policy space Γ is learnable, although they are silent about rates of convergence. Taken altogether, we believe our two sets of theoretical results provide a comprehensive means of making and evaluating policy decisions.

This paper also makes a contribution from an identification perspective. Perhaps unsurprisingly, an important theoretical object in our study of policy decisions are the following policy transform envelope functions:

$$I_{\ell b}[\varphi](\gamma) := \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s), \qquad I_{ub}[\varphi](\gamma) := \sup_{s \in \mathcal{S}} I[\varphi](\gamma, s).$$

Regardless of the true (sub-)state $s_0 \in \mathcal{S}$, at the true distribution $P_{Y,Z}$ the policy transform in (3.1) can be "sandwiched" between these upper and lower envelope functions. This idea is illustrated in Figure 3.1. Our ability to provide a tractable characterization of these envelope functions thus turns out to be critical to our ability to provide sufficient conditions for policy learnability, and for our ex-post analysis of the eME rule and the $\delta-$level set.

The envelope functions may not be policy transforms themselves, but under some conditions they can be interpreted as sharp bounds on the policy transform $I[\varphi](\gamma)$, point-wise in the variable γ. It is here that we make a contribution in the identification literature by showing that the envelope functions can be expressed as the value functions of optimization problems parameterized by the policy variable $\gamma \in \Gamma$. The result is derived under assumptions found in the theory of error bounds and exact penalty functions from the literature on optimization, and the resulting optimization problems are closely related to *mathematical*

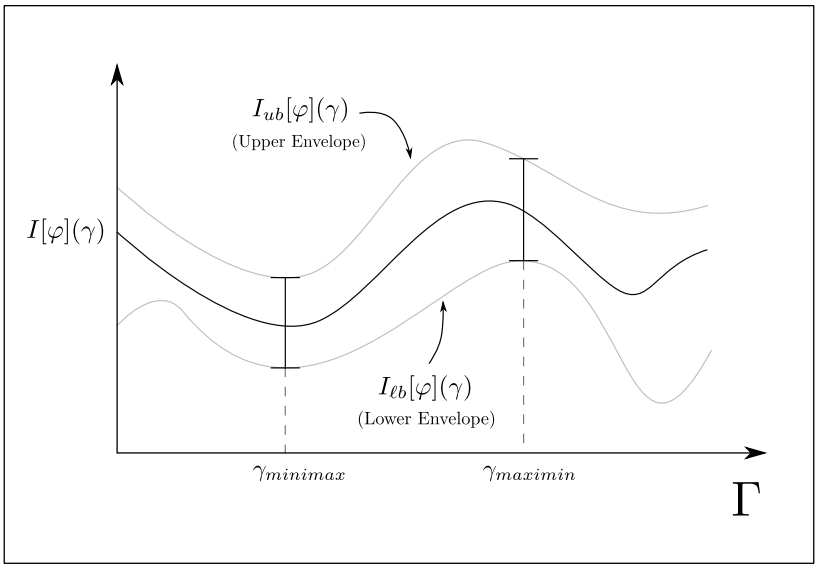


Figure 3.1: This figure illustrates the policy transform of some function φ, as well as the upper and lower envelope functions $I_{ub}[\varphi](\gamma)$ and $I_{\ell b}[\varphi](\gamma)$ (resp.). The minimax (over (sub-)states $s \in \mathcal{S}$) policy is the policy that minimizes the upper envelope, and the maximin (over (sub-)states $s \in \mathcal{S}$) policy choice is the policy that maximizes the lower envelope.

programs with equilibrium constraints, or MPECs.[4] A remarkable benefit of our optimization approach is that we show the bounds on the policy transform can be constructed without the need to first estimate the full identified set for θ, the vector of model parameters. This is in contrast to typical approaches to bounding counterfactual quantities, which first estimate the identified set of structural parameters, and then perform a counterfactual for every possible value of the parameter vector in the identified set. A direct implication of our result is that, in either point- or partially-identified models, if the policymaker's counterfactual quantity of interest is the policy transform of some function φ, then all structural parameters can be treated as nuisance parameters when performing counterfactuals and making policy choices. These results on identification may be of substantial separate interest.

Finally, throughout the text we discuss a simultaneous discrete choice and a program evaluation example in order to illustrate possible applications of the procedure. The simultaneous discrete choice example includes empirical entry games (e.g. Tamer (2003)) and empirical models of social interactions (e.g. Brock and Durlauf (2001)) as special cases, and has become a canonical example of an incomplete model in the literature on partial identification. The second program evaluation example follows the setup in Heckman and Vytlacil (1999) and Heckman and Vytlacil (2005). This example has attracted recent attention in the literature on partial identification (e.g. Mogstad et al. (2018) and Mourifie and Wan (2020)) and is included to show the breadth of our procedure.

3.1.1 Related Literatures

This paper builds on results from a variety of different literatures, including recent work on counterfactuals in structural models, partial identification and random set theory, decision theory and optimal policy choice, and computational and statistical learning theory.

Our approach to modelling and counterfactuals in partially identified models extends the literature using random set theory in econometrics, including Beresteanu et al. (2011), Galichon and Henry (2011),

[4]See Dolgopolik (2016) for a survey of exact penalty functions and their connection to error bounds, and see Luo et al. (1996) for a textbook treatment MPECs.

Beresteanu et al. (2012) and Chesher and Rosen (2017a). As mentioned in the introduction, our general environment is similar to descriptions found in Jovanovic (1989) and more recently in Chesher and Rosen (2017a), which in turn are extensions of the classical foundations for econometric modelling set forth in Koopmans et al. (1950) and Hurwicz (1950), among others. The use of random set theory is convenient in order to permit application of the method to a wider class of models. In particular, our framework is applicable to models that may (or may not) be incomplete, which are an important class of models in the literature on partial identification. Incomplete models are now legion, and include entry games with multiple equilibria (Bresnahan and Reiss (1990), Bresnahan and Reiss (1991), Tamer (2003), Jia (2008), Ciliberto et al. (2018)); english auctions (Haile and Tamer (2003), Chesher and Rosen (2017b)); discrete choice models with endogenous regressors or social interactions (Chesher and Rosen (2012), Chesher et al. (2013), Chesher and Rosen (2014)); matching models (Uetake and Watanabe (2019)); friendship networks (Miyauchi (2016)); and selection and treatment effect models (Mourifie et al. (2018), Russell (2019)).

From the perspective of policy choice, our general approach to the problem of policy decisions is new. However, there is now a large and growing literature on statistical treatment rules in econometrics, including papers by Manski (2004), Hirano and Porter (2009), Stoye (2009a), Stoye (2012), Chamberlain (2011), Tetenov (2012), Kasy (2016), Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2019). In general these papers can be divided according to (i) whether they are frequentist/bayesian, (ii) whether they take a finite-sample or asymptotic approach, and (iii) whether they consider decision problems under uncertainty or ambiguity (or "Knightian uncertainty"). In the current paper we take a frequentist, finite-sample approach to decision problems under ambiguity. However, unlike previous papers that belong to the same class, our method of evaluating statistical decision rules differs from the procedure proposed by Wald (1950). In the absence of ambiguity arising from the unknown sub-state $s \in \mathcal{S}$, our procedure is very similar to the PAC framework for inductive inference that has become enormously popular in the computer science literature. This model of learning was initially proposed in a seminal paper by Valiant (1984), for which he won the prestigious Turing Award. The name "probably approximately correct" seems to have been first used by Angluin and Laird (1988), who extended the model to the case of noisy data. The PAC model and its extensions have now become the dominant model of learning in the theoretical foundations of machine learning; influential textbook treatments that make this connection explicit include Kearns et al. (1994), Vapnik (1995), Vapnik (1998), Vidyasagar (2002), Shalev-Shwartz and Ben-David (2014) and Mohri et al. (2018). Our particular analysis is most closely related to the decision theoretic generalization of the PAC learning model proposed by Haussler (1992), as well as the general learning setting considered in Vapnik (1995). Other important papers studying necessary and sufficient conditions for learnability in various machine learning settings include Blumer et al. (1989), Kearns and Schapire (1994), Bartlett et al. (1996), Alon et al. (1997), and Shalev-Shwartz et al. (2010), among others. Our work here on providing sufficient conditions for learnability borrows heavily from this literature. However, the additional ambiguity that arises in relation to possible partial identification of the policy transform differentiates our setting from the statistical learning literature, and our incorporation of this notion of ambiguity into the PAC framework appears to be new. Many of our results are applicable to problems involving risk minimization subject to (stochastic) constraints, and thus may be of separate interest to researchers in machine learning.

Surprisingly, we are unaware of any attempts to formally connect the literature on statistical decision theory with the literature on statistical learning theory.[5] On the one hand, the properties of a Wald-style analysis are (at this point) better understood; see, for example, Stoye (2011) for an axiomatization of Wald's

[5] Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2019) make some connections with the statistical learning literature. However, their method of evaluating statistical treatment rules is different from that considered by the PAC model. Some discussion on the links with decision theory can be found in an influential paper by Haussler (1992), although the discussion is very limited and no connection is made with Wald-style frequentist decision theory. As far as we are aware, this remains an open question. We make a preliminary comparison in Appendix 3.A.3.

frequentist maximin procedure. On the other hand, we find the PAC style criterion to be much more amenable to informative ex-post analyses of particular decision rules, mostly due to its connection to the concentration of measure phenomenon, and thus its amenability to analysis using concentration inequalities.

The connections to the statistical learning literature permeate our theoretical results. There are connections of our work to the study of ratio-type empirical processes (e.g. Giné et al. (2003), Giné et al. (2006)), and to the study of fixed-point equations and rates of convergence in risk minimization problems (e.g. Massart (2000), Koltchinskii and Panchenko (2000), Bousquet et al. (2002), Bartlett et al. (2005), and Koltchinskii (2006)). Overall our work is most closely related to the work of Koltchinskii (2006), and the subsequent textbook treatment Koltchinskii (2011). As we will see in the section on the ex-post analysis of certain decision rules, a key component of our approach is the use of Rademacher processes to construct data-dependent bounds on certain important empirical processes. This has the benefit of allowing the policymaker to avoid relying on any specific properties of the underlying function class, which are typically difficult to verify, and thus are applicable whether or not the associated policy space is learnable. Furthermore, the use of data-dependent complexity measures like the empirical Rademacher complexity ensures our finite sample guarantees are less conservative than otherwise. It appears this idea was independently offered by Bartlett et al. (2002) and Koltchinskii (2001), and was developed further in Koltchinskii (2006). See also Section 4.2 in Koltchinskii (2011). A review of excess risk bounds and their application to classification problems in statistical learning theory can be found in Boucheron et al. (2005) and Koltchinskii (2011).

Closely related to the identification component of this paper—which studies the envelope functions for the policy transform—is the work by Ekeland et al. (2010), Schennach (2014), Torgovitsky (2019) and Li (2019). The paper of Ekeland et al. (2010) is focused on model specification testing, and allows for econometric models with only semiparametric restrictions on the distribution of unobservables in the form of moment conditions.[6] Schennach (2014) provides a general framework for models with moment conditions that depend on latent variables, and shows that the latent variables can be integrated out of the moment conditions without loss of information using a least-favourable entropy maximizing distribution. Torgovitsky (2019) shows that when restrictions on the distribution of the latent variables have a certain structure, sharp identified sets for functionals of partially-identified parameters can be characterized in terms of optimization problems. Finally, Li (2019) shows that sharp identified sets for structural and counterfactual parameters can be constructed using a method that essentially profiles the latent variables out of the moment conditions. In the current paper, we use an idea related to Li (2019) to eliminate unobservables from the counterfactual bounding problem. However, in contrast to Li (2019) our focus on policy transforms means our formulation does not require replacing a finite number of moment conditions with a continuum of moment conditions. Furthermore, our approach does not require the policymaker to compute the full identified set of structural parameters. Our specific characterization of the bounds on the policy transform in terms of two parametric optimization problems was designed to be amenable to the theoretical analysis of policy space learnability, and the analysis of the eME rule and the $\delta-$level sets. Thus, our particular bounding approach is new. Finally, and perhaps most importantly, our focus is primarily on using the bounds to study the problem of policy choice, which is not considered in any of Ekeland et al. (2010), Schennach (2014), Torgovitsky (2019) or Li (2019).

The idea that at least some structural parameters may be seen as nuisance parameters in the policy decision problem goes back at least as far as Marshak (1953). Heckman (2010) refers to this idea as "Marshak's Maxim." At a high level, the identification component of this paper is reminiscent of Ichimura and Taber (2000), who discuss a method for performing ex-ante policy experiments in the treatment effect literature without estimating the structural parameters, and without specifying the error distribution. More recent

[6] The paper of Ekeland et al. (2010) is related to a string of other papers by the same authors, namely Galichon and Henry (2006), Galichon and Henry (2009) and Galichon and Henry (2011).

examples of counterfactual analysis without first estimating the (identified set for the) structural parameters can be found in Syrgkanis et al. (2018), Tebaldi et al. (2019) and Kalouptsidi et al. (2019).

The remainder of the paper will proceed as follows. Section 3.2 introduces the notation and main definitions and assumptions, in addition to describing the decision environment and introducing the motivating examples. Importantly, Section 3.2 introduces the policy transform, and defines the notion of learnability of a policy space. As described above, the theoretical results in this paper depend heavily on the nature of the upper and lower envelope functions for the policy transform. Thus, in Section 3.3 we define the identified set for the policy transform, and present our main identification result characterizing its upper and lower envelopes. Equipped with this result, Section 3.4 then considers the problem of policy choice, providing sufficient conditions for learnability, and Section 3.5 provides an ex-post analysis of the performance of particular decision rules. Section 3.6 concludes. All proofs can be found in the Appendices.

Notation: Given a subset $\mathcal{X}$ of a Polish space (a complete and separable metric space), we use $\mathfrak{B}(\mathcal{X})$ to denote the Borel $\sigma-$algebra on $\mathcal{X}$ (note the topology on $\mathcal{X}$ is the topology induced by the metric). We will often either leave the metric implicit, or will denote a generic metric by the function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For two measurable spaces $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ and $(\mathcal{X}', \mathfrak{B}(\mathcal{X}'))$, the product $\sigma-$algebra on $\mathcal{X} \times \mathcal{X}'$ is denoted by $\mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{X}')$. If $X : (\Omega, \mathfrak{A}) \to (\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is a random variable defined on the probability space $(\Omega, \mathfrak{A}, P)$, then we use P_X to denote the probability measure induced on $\mathcal{X}$ by X; that is, for any $A \in \mathfrak{B}(\mathcal{X})$, $P_X(A) := P(X^{-1}(A))$. We let $\sigma(X) \subseteq \mathfrak{A}$ denote the smallest sub $\sigma-$algebra making X a measurable function. Furthermore, we interpret $P_{X|X'}(X \in A|X' = x)$ as a regular conditional probability measure. In many cases we do not explicitly differentiate between the true distribution of the random variable X, say P_X, or some other distribution of the random variable X, say P'_X, and instead leave the distinction to be resolved by context. To keep the notation clean, we will omit the transpose when combining column vectors; that is, if v_1 and v_2 are two column vectors, rather than write $v = (v_1^\top, v_2^\top)^\top$ we instead write $v = (v_1, v_2)$, where it is understood that v is a column vector unless otherwise specified. Importantly, throughout the paper we use the convention that $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$. Finally, we will largely ignore measurability issues in the main text, but we note that such issues are non-trivial in our framework, and are discussed and addressed in Appendix 3.B.2.

3.2 Methodology

3.2.1 Preliminaries

As mentioned in the introduction, the description of the environment follows closely that of Jovanovic (1989) and Chesher and Rosen (2017a), which in turn are extensions of the classical foundations for econometric modelling set forth in Koopmans et al. (1950) and Hurwicz (1950), among others. However, there are some differences that will be pointed out as they occur. We will also make heavy use of random set theory in this paper. Random set theory has played a major role in the development of methods for partially identified models, for example in the contributions of Beresteanu et al. (2011), Galichon and Henry (2011), Beresteanu et al. (2012) and Chesher and Rosen (2017a), among others. We will also use random set theory in this paper, as it naturally generalizes many features of complete econometric models to incomplete models (see Chesher and Rosen (2017a)). Since complete models can be seen as special cases of incomplete models, focusing on incomplete models will allow us to construct a method that applies to a broader class of econometric models. Some important definitions from random set theory—including the notion of Effros-measurability,

the definition of a random set, the distribution of a random set, and the notion of a selection from a random set—have been moved to Appendix 3.A for brevity. The current section will presume some working knowledge of these concepts.

We begin by specifying the restrictions on the factual and counterfactual domains. First we will fix the probability space and define the unobserved random variables and parameters that are common to both domains.

Assumption 3.2.1. *There exists a fixed probability space* $(\Omega, \mathfrak{A}, P)$*, and a random element* $U : (\Omega, \mathfrak{A}) \to (\mathcal{U}, \mathfrak{B}(\mathcal{U}))$ *where* $\mathcal{U}$ *is a compact second-countable Hausdorff space. In addition, the parameter space* Θ *is a Polish space equipped with the* $\sigma-$*algebra* $\mathfrak{B}(\Theta)$*.*

Fixing the probability space throughout represents a departure from some of the existing literature on partial identification and random set theory in econometrics (e.g. Galichon and Henry (2011), Chesher and Rosen (2017a)). Our reason for doing so is mostly conceptual. This paper is concerned with counterfactuals, and counterfactuals naturally involve some comparison of units between factual and counterfactual states. In any probabilistic framework, the underlying probability space naturally specifies the basic unit of observation (e.g. individuals, firms, types, etc.), so that it is necessary for the units of observation to be the same in both the factual and counterfactual states when performing a counterfactual analysis. The point may seem esoteric, but it will have a major impact on the statement and proofs of most of our results while also resolving some interpretative difficulties.

The restriction that $\mathcal{U}$ is a compact space in Assumption 3.2.1 may seem overly restrictive; for example, the euclidean space $\mathbb{R}^d$ $(d < \infty)$ with the usual topology is not a compact space. We might consider relaxing Assumption 3.2.1 by allowing $\mathcal{U}$ to be a locally compact second-countable Hausdorff space, of which $\mathbb{R}^d$ (with the usual topology) is an example. However, any locally compact Hausdorff space has a one-point compactification; that is, assuming $\mathcal{U}$ is locally compact and Hausdorff, there exists a compact space $\widetilde{\mathcal{U}}$ with $\mathcal{U} \subset \widetilde{\mathcal{U}}$ such that $\widetilde{\mathcal{U}} \setminus \mathcal{U}$ consists of a single point.[7] Furthermore, $\widetilde{\mathcal{U}}$ is unique up to a homeomorphism.[8] A related argument has been presented in Schennach (2014). From this perspective, it is difficult to imagine an environment where a policymaker should have strong a priori reasons to model the unobservables using a locally compact Hausdorff space $\mathcal{U}$ versus its one-point compactification $\widetilde{\mathcal{U}}$, despite the fact that this is often done in practice. On the other hand, the theoretical benefits of taking $\mathcal{U}$ to be compact (or to be the one-point compactification of some locally compact Hausdorff space) are numerous. We will highlight these benefits as they arise.

Note that we will not require that the distribution U belong to a parametric class. This is in keeping with our desire to avoid treating the distribution of U as a model primitive. This perspective is consistent with the idea that the latent variables represent components of the underlying economic system that remain unmodelled, due primarily to the policymaker's ignorance of the process determining U, and thus her inability to construct a complete mathematical description of the economic system under investigation. This interpretation becomes especially meaningful given the role the latent variables play in determining counterfactual outcomes. Instead, as we will see, the distribution of U can be implicitly constrained by the remaining primitives of the model.

Finally we note that equipping the parameter space with the Borel $\sigma-$algebra $\mathfrak{B}(\Theta)$ may seem odd. However, to make policy decisions in our framework will require measurability of certain functions to be introduced later on. Primitive conditions for the required measurability will make use of the measure space $(\Theta, \mathfrak{B}(\Theta))$. We return to similar points throughout the paper, and refer readers to Appendix 3.B.2 for our results on measurability.

[7] See Munkres (2014) Theorem 29.1.
[8] Recall a homeomorphism is a continuous invertible function with a continuous inverse.

We will now summarize the restrictions on the factual and counterfactual domains, beginning with the factual domain.

Assumption 3.2.2 (Factual Domain)**.** *The factual domain is represented by random vectors* $Y : (\Omega, \mathfrak{A}) \to (\mathcal{Y}, \mathfrak{B}(\mathcal{Y}))$ *and* $Z : (\Omega, \mathfrak{A}) \to (\mathcal{Z}, \mathfrak{B}(\mathcal{Z}))$*, where* $\mathcal{Y}$ *and* $\mathcal{Z}$ *are Polish spaces. There exists a (possibly multi-valued) map* $\boldsymbol{G}^- : \mathcal{Y} \times \mathcal{Z} \times \Theta \to \mathcal{U}$ *which is closed and Effros-measurable, and satisfies:*

$$P\left(U \in \boldsymbol{G}^-(Y, Z, \theta_0) | Y = y, Z = z\right) = 1, \tag{3.4}$$

$(y, z)-$*a.s. for some* $\theta_0 \in \Theta$*. Furthermore,*

$$\mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}}[m_j(y, z, u, \theta_0)] \leq 0, \qquad j = 1, \ldots, J, \tag{3.5}$$

for some measurable functions $m_j : \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \Theta \to \mathbb{R}$*, for* $j = 1, \ldots, J$*, bounded in absolute value for each* $\theta \in \Theta$*.*

The first part of the assumption states that the unobserved random vector is a selection from the random set $\boldsymbol{G}^-(Y, Z, \theta_0)$ (see Appendix 3.A for the definition of a selection).[9] Note the assumption requires only that $\boldsymbol{G}^-(\cdot, \theta)$ admits a selection when $\theta = \theta_0$. The first part of the assumption can thus be interpreted as a support restriction for the vector of unobservables conditional on the observed data. These support restrictions are derived from the policymaker's econometric model, as we will see in the examples ahead. We also note that the random set $\boldsymbol{G}^-$ contains the $U-$level sets presented in Chesher and Rosen (2017a) as a special case, and thus our framework will be applicable to the *generalized instrumental variable* (GIV) models considered in their work.

In the second part of the assumption we suppose that the factual domain satisfies the moment inequalities in (3.5), which are allowed to depend on the unobserved random variable U. This differs from moment conditions in the generalized method of moments (GMM), as well as typical definition of moment inequalities (c.f. Chernozhukov et al. (2007b)). This places our paper in the narrow literature in partial identification that allows for moments to depend on unobserved random variables with a possibly unknown distribution (c.f. Ekeland et al. (2010), Schennach (2014), Torgovitsky (2019) and Li (2019)). The assumption of boundedness of the moment functions may appear to be restrictive. This assumption might be replaced by the weaker assumption that the moment functions are uniformly integrable with respect to the set of probability measures $P_{U|Y,Z} \times P_{Y,Z}$ satisfying the other components of Assumption 3.2.2.[10] However, regardless of how it is weakened, we contend that boundedness of the moment functions remains the most primitive assumption for our purposes. Finally, the fact that there are only a finite number of moment functions may also be restrictive; for example, this prohibits the use of conditional moment inequalities when the conditioning variable is continuous. Our identification result in Section 3.3 can be extended—under a suitable modification of our assumptions—to handle the case of an infinite number of moment inequalities. However, the same statement is not true of the results in Sections 3.4 and 3.5 on policy decisions, which rely more crucially on the fact that the number of moment conditions is finite. We also note that both the Effros measurability of $\boldsymbol{G}^-$ and Borel measurability of each moment function m_j with respect to $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\Theta)$ (rather than only with respect to $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z})$) will be required later on to ensure measurability of certain key classes of functions.

Similar to the factual domain, we must specify restrictions on the counterfactual domain, and when specifying the counterfactual domain we must specify which counterfactuals are under consideration by

[9] A similar argument to the one presented in Appendix B of Chesher and Rosen (2015) can be used to show that this characterization of selectionability conditional on (y, z) a.s. is equivalent to using an analogous selectionability criteria for the joint distributions of (Y, Z, U). A similar point will apply later on when we introduce Assumption 3.2.3.

[10] See for example alternative assumptions given in Ekeland et al. (2010) and Li (2019).

the policymaker. We index various counterfactuals by an abstract parameter γ, where a fixed value of γ represents a single counterfactual, and different values of γ correspond to different counterfactuals. The interpretation of the parameter γ that will be used throughout is that it is an abstraction of a policy tool under the control of the policymaker. The parameter γ will play an important role in our policy decision procedure presented later in the paper.

Assumption 3.2.3 (Γ-Counterfactual Domains)**.** *The Γ-counterfactual domains are represented by a stochastic process $\{Y^\star(\omega,\gamma) : \gamma \in \Gamma\}$ where $(\Gamma, \mathfrak{B}(\Gamma))$ is a measurable space with Γ a Polish space, and where $Y^\star_\gamma := Y^\star(\cdot,\gamma)$ is such that $Y^\star : (\Omega \times \Gamma, \mathfrak{A} \otimes \mathfrak{B}(\Gamma)) \to (\mathcal{Y}^\star, \mathfrak{B}(\mathcal{Y}^\star))$ is measurable, with $\mathcal{Y}^\star$ a Polish space. Furthermore, there exists a (possibly multi-valued) map $\boldsymbol{G}^\star : \mathcal{Y} \times \mathcal{Z} \times \mathcal{U} \times \Theta \times \Gamma \to \mathcal{Y}^\star$ which is closed and Effros measurable, and satisfies:*

$$P\left(Y^\star_\gamma \in \boldsymbol{G}^\star(Y,Z,U,\theta_0,\gamma) | Y=y, Z=z, U=u\right) = 1, \tag{3.6}$$

$(y,z,u)-$a.s. for the same $\theta_0 \in \Theta$ from Assumption 3.2.2, and for all $\gamma \in \Gamma$.

Compared to the existing literature, Assumption 3.2.3 appears to be new. It restricts the set of counterfactuals considered in this paper to be those that can be written as modifications of support-like restrictions on the random variables in the model. We contend that this assumption is able to accommodate most counterfactuals of interest in economics, although it rules out, for example, consideration of counterfactuals that modify the distributions of the latent variables. Under this assumption we have that $Y^\star_\gamma := Y^\star(\cdot,\gamma)$ is a *selection process* from the *set-valued process* $\boldsymbol{G}^\star(Y,Z,U,\theta_0,\gamma)$, where $\boldsymbol{G}^\star$ is required to be Effros-measurable with respect to the product $\sigma-$algebra. Again, the measurability requirement with respect to both Θ and Γ may seem odd, but will be required in Section 3.4 and 3.5 when we consider the question of policy choice. Note that—consistent with the remark following Assumption 3.2.1—the probability space in Assumptions 3.2.2 and 3.2.3 are assumed to be the same.

Remark 3.2.1 (The "No Back-Tracking" Principle)**.** *From a purely mathematical standpoint there is no reason that the moment functions in Assumption 3.2.2 cannot also be functions of $Y^\star_\gamma$ and/or $\gamma \in \Gamma$. However, we omit this extension for interpretive reasons and caution researchers interested in this approach. In particular, if the researcher is not judicious in her formulation of such moment functions, then it is possible to have environments where the counterfactual $\gamma \in \Gamma$ of interest has "identifying power" for the structural parameters $\theta \in \Theta$. Such environments are extremely puzzling since, intuitively, in these cases the counterfactual domain $\gamma \in \Gamma$ under consideration contains "information" on the values of the structural parameters $\theta \in \Theta$ existing in the factual domain. Environments that avoid such difficulties will be said to satisfy the "no back-tracking principle."*[11] *We will return to this idea at some point in our example on simultaneous discrete choice models.*

The setup implied by Assumptions 3.2.1, 3.2.2 and 3.2.3 is illustrated in Figure 3.2. Throughout the remainder of the paper, we let $V_\gamma := (Y^\star_\gamma, Y, Z, U)$ denote a random vector with realizations $v \in \mathcal{V}$, where $\mathcal{V}$ is a product space with the product $\sigma-$algebra.

3.2.2 Examples

We will now turn to two examples to help illustrate the nature of the assumptions just introduced. The examples will be revisited throughout the remainder of the text. The introduction of the examples is lengthy, and readers may skip to Subsection 3.2.3 without loss of continuity.

[11] This principle is named in honour of the philosopher David Lewis who argued against similar "back-tracking counterfactuals" in Lewis (1979).

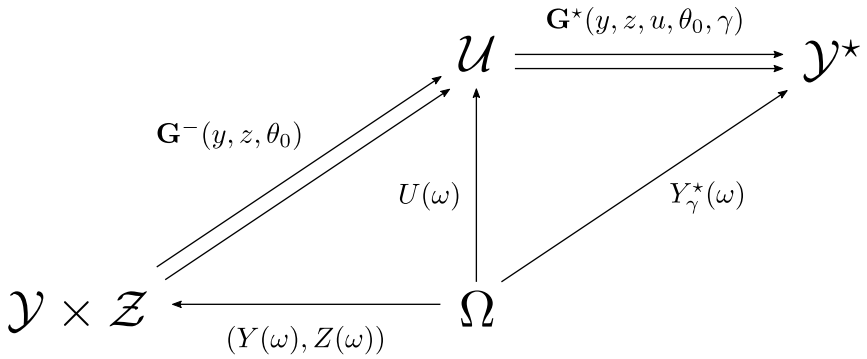


Figure 3.2: Displayed above is an illustration of the setup implied by Assumptions 3.2.1, 3.2.2 and 3.2.3. In particular, note that all random variables are assumed to be defined on the same probability space. Furthermore, note the direction of the arrows from the factual domain $\mathcal{Y} \times \mathcal{Z}$ to the latent $\mathcal{U}$ to the counterfactual domain $\mathcal{Y}^{\star}$, intended to illustrate the process by which information from the factual domain informs on the counterfactual domain.

The first example we consider is a simultaneous discrete choice model. Simultaneous discrete choice models have seen a wide number of applications, including empirical entry games (e.g. Tamer (2003)), and discrete choice models with social interactions (e.g. Brock and Durlauf (2001)). It is already known from the work of Chesher and Rosen (2020) that this example falls into the class of GIV models considered by Chesher and Rosen (2017a). For readers familiar with these works, the model will serve as a natural point of comparison. The second example is a program evaluation example which closely mirrors the environment in Heckman and Vytlacil (2005). The example shows a model where the structural parameter is point-identified, but the counterfactual object of interest is partially-identified.[12]

Example 1 (Simultaneous Discrete Choice)**.** *Consider a simultaneous discrete choice problem. In particular, assume that a binary outcome vector* $Y := (Y_1, \ldots, Y_K)$ *has generic element* $Y_k \in \mathcal{Y}$ *determined by the equation:*

$$Y_k = \mathbb{1}\{\pi_k (Z_k, Y_{-k}; \theta) \geq U_k\}. \tag{3.7}$$

Here Z_k *is a vector of covariates,* U_k *is an unobserved random variable, and* θ *is a vector of model parameters. We will define the vector* $Z := (Z_1, \ldots, Z_K)$ *and* $U := (U_1, \ldots, U_K)$ *where each variable* Z_k *has support* $\mathcal{Z} = \{z_1, \ldots, z_L\}$*, a finite subset of euclidean space, and each* U_k *has support* $\mathcal{U} = [-1, 1]^{d_u}$*.*[13] *For each* k*, we assume that* π_k *is a known measurable function of* (Z_k, Y_{-k}, θ)*, mapping to* $[-1, 1]$ *that is linear in parameters* θ *and has a gradient (with respect to* θ*) bounded away from zero for each* (z, y_{-k})*. We also assume that* $\theta = (\theta_1, \ldots, \theta_K)$*, and that each* π_k *depends only on the subvector* θ_k*. For simplicity we will assume that the parameter space* Θ *is a compact subset of* $\mathbb{R}^{d_\theta}$*, and that* U *is continuously distributed. To illustrate the use of semi-parametric restrictions, we will also assume that each coordinate of the vector* U *is (i) median zero, and (ii) median independent of* (Z_k, Y_{-k})*. Finally, we assume all random variables are supported on the same probability space* $(\Omega, \mathfrak{A}, P)$*. Verification of Assumption 3.2.1 under these conditions is presented in Appendix 3.C.1.*

[12] In our setting, this is due to the fact the instrument will be assumed to have finite support.

[13] Note that we could instead define $\mathcal{U} := \overline{\mathbb{R}}^{d_u}$, but then:

$$\mathbb{1}\{\pi_k (Z_k, Y_{-k}; \theta) \geq U_k\} = \mathbb{1}\{\tilde{\pi}_k (Z_k, Y_{-k}; \theta) \geq \tilde{U}_k\},$$

where $\tilde{\pi}_k (Z_k, Y_{-k}; \theta) = \tanh(\pi_k (Z_k, Y_{-k}; \theta))$ and $\tilde{U}_k = \tanh(U_k)$. In other words, the case with $\mathcal{U} := \overline{\mathbb{R}}^{d_u}$ is homeomorphic to the case $\mathcal{U} := [-1, 1]^{d_u}$.

For the factual domain, we have the following multifunction:

$$\boldsymbol{G}^{-}(Y,Z,\theta) := cl\{u \in \mathcal{U} : Y_k = \mathbb{1}\{\pi_k(Z_k, Y_{-k}; \theta) \geq u_k\},\ k = 1,\ldots,K.\}. \tag{3.8}$$

Note the closure is taken to ensure that $\boldsymbol{G}^{-}(\cdot,\theta)$ is a closed set for each θ. However, this introduces no additional structure and serves merely as a technical simplification, since $\boldsymbol{G}^{-}(\cdot,\theta_0)$ as defined above will be almost surely equal to the right hand side of (3.8) without taking the closure, which follows from the assumption that U is continuously distributed. To complete the description of the factual domain, we will impose the median zero and median independence assumptions for each coordinate of the vector U as a sequence of moment conditions. In particular, for $k = 1\ldots,K$, we will impose the moment conditions:

$$\mathbb{E}[(\mathbb{1}\{U_k \geq 0\} - \mathbb{1}\{U_k \leq 0\})\,\mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}] \leq 0, \quad \forall z \in \mathcal{Z},\ y_{-k} \in \mathcal{Y}^{K-1}, \tag{3.9}$$

$$\mathbb{E}[(\mathbb{1}\{U_k \leq 0\} - \mathbb{1}\{U_k \geq 0\})\,\mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}] \leq 0, \quad \forall z \in \mathcal{Z},\ y_{-k} \in \mathcal{Y}^{K-1}. \tag{3.10}$$

Taken together, (3.9) and (3.10) imply that the latent variables U_k are both median zero and median independent of covariates Z_k and the outcomes Y_{-k}.[14] *Verification of Assumption 3.2.2, including Effros-measurability of the multifunction (3.8), is provided in Appendix 3.C.1.*

Turning to the counterfactual domain, there are many possible counterfactuals that may be of interest. For the sake of illustration, we will consider counterfactuals of the following form. Let $\gamma_k : \mathcal{Z} \times \mathcal{Y}^{K-1} \to \mathcal{Z} \times \mathcal{Y}^{K-1}$, $\gamma = (\gamma_k)_{k=1}^{K}$, and $Y^{\star}_{\gamma} := (Y^{\star}_{1,\gamma},\ldots,Y^{\star}_{K,\gamma})$ with typical element:

$$Y^{\star}_{k,\gamma} = \mathbb{1}\{\pi_k(\gamma(Z_k, Y^{\star}_{-k,\gamma}); \theta) \geq U_k\}. \tag{3.11}$$

For example, our interest may be in the properties of the counterfactual random variable $Y^{\star}_{k,\gamma}$, such as its mean or its conditional mean. The multifunction for the counterfactual domain is then given by:

$$\boldsymbol{G}^{\star}(Z,U,\theta,\gamma) := \left\{y^{\star} \in \mathcal{Y} : y^{\star}_k = \mathbb{1}\{\pi_k\left(\gamma(Z_k, y^{\star}_{-k}); \theta\right) \geq U_k\},\ k = 1,\ldots,K.\right\}. \tag{3.12}$$

Note here we take $\mathcal{Y}^{\star} = \mathcal{Y}$. Verification of Assumption 3.2.3, including Effros-measurability of the multifunction in (3.12), is provided in Appendix 3.C.1.

Example 2 (Program Evaluation)**.** *Consider the problem of program evaluation. In this example, a binary variable $D \in \{0,1\}$ indicates participation in the treatment or control group for some program, and the observed real-valued outcome is given by:*

$$Y = U_0(1-D) + U_1 D, \tag{3.13}$$

where U_0 and U_1 are potential outcomes that are never jointly observed. We will assume throughout that $U_0, U_1 \in \mathcal{U} = [\underline{Y}, \overline{Y}]$, and thus we also assume the outcome Y takes values in the bounded interval $\mathcal{Y} := [\underline{Y}, \overline{Y}]$. In the absence of a selection equation determining the values of D, the potential outcome model is incomplete. This case is considered in Russell (2019), and the framework in this paper applies to this case as well. Alternatively, we will consider the more popular approach of Heckman and Vytlacil (1999) and Heckman and Vytlacil (2005), and will suppose that the treatment is determined by the equation:

$$D = \mathbb{1}\{g_0(Z) \geq U\}, \tag{3.14}$$

[14] Note that this restriction implies constraints on the joint distribution of the vector $(U_1,\ldots,U_K)$. Alternatively, we might instead impose only median independence of U_k with Z_k, which restricts only the marginal distribution of U_k.

where U is continuous, and $g_0(\cdot)$ is an unknown measurable function of the observable covariates $Z \in \mathcal{Z} \subset \mathbb{R}^{d_z}$, where d_z is the dimension of the vector Z. We will assume that $\mathcal{Z}$ is finite, and will allow for the case when the vector Z can be decomposed as $Z = (X, Z_0)$ with (i) $U \perp\!\!\!\perp Z_0|X$ (conditional independence) and (ii) $\mathbb{E}[U_d|Z] = \mathbb{E}[U_d|X]$ for $d \in \{0,1\}$ (mean independence). We will thus decompose $\mathcal{Z}$ as $\mathcal{Z} = \mathcal{Z}_0 \times \mathcal{X}$, where $\mathcal{Z}_0$ is the support of Z_0 and $\mathcal{X}$ is the support of X. Under these assumptions, it is without loss of generality that U be taken to be uniformly distributed on $[0,1]$ conditional on Z. As shown in Vytlacil (2002), these assumptions, combined with the additive separability of the selection equation in (3.14), are equivalent to the assumptions required to estimate the local average treatment effect (LATE) of Imbens and Angrist (1994). This connects this model with a large body of empirical work that focuses on obtaining estimates of the LATE.

Set the parameter space as $\Theta = \mathcal{G} \times \mathcal{T}$. Here $\mathcal{G}$ can be taken equal to the space of all positive measurable functions on $\mathcal{Z}$ which is a metric space with the sup norm (for example); under finiteness of $\mathcal{Z}$, this space is Polish. Furthermore, we will take the component $\mathcal{T}$ of the parameter space to be the space of all possible measurable functions on $\mathcal{Z}$. This component of the parameter space will be used in the moment functions below. Finally, we shall denote a generic pair $(g,t) \in \Theta$ as θ.

We will denote the support of (U_0, U_1, U) as $\mathcal{U} := [\underline{Y}, \overline{Y}]^2 \times [0,1]$. We also assume that the random variables in the vector (Y, D, Z, U_0, U_1, U) are all supported on the same probability space $(\Omega, \mathfrak{A}, P)$. Under these conditions, Assumption 3.2.1 is verified in Appendix 3.C.2.

For the factual domain we have the multifunction:

$$\boldsymbol{G}^-(Y, D, Z, \theta) := cl\left\{(U_0, U_1, U) \in \mathcal{U} : \begin{array}{l} Y = U_0(1-D) + U_1 D, \\ D = \mathbb{1}\{g(Z) \geq U\}. \end{array}\right\}. \tag{3.15}$$

Note the closure is taken to ensure that $\boldsymbol{G}^-(\cdot, \theta)$ is a closed set for each θ. However, this introduces no additional structure and serves merely as a technical simplification, since $\boldsymbol{G}^-(\cdot, \theta)$ as defined above will be almost surely equal to the right hand side of (3.15) without taking the closure, which follows from the assumption that U is continuously distributed. Close inspection of this multifunction provides some simplification:

$$\boldsymbol{G}^-(Y, D, Z, \theta) = \begin{cases} \{Y\} \times [\underline{Y}, \overline{Y}] \times [g(Z), 1], & \text{if } D = 0, \\ [\underline{Y}, \overline{Y}] \times \{Y\} \times [0, g(Z)], & \text{if } D = 1. \end{cases} \tag{3.16}$$

To complete the description of the factual domain, we will impose the independence condition $U \perp\!\!\!\perp Z_0|X$ and the mean independence condition $\mathbb{E}[U_d|Z] = \mathbb{E}[U_d|X]$, for $d \in \{0,1\}$, as a sequence of moment conditions. In particular, since $\mathcal{Z}$ is assumed to be finite, let us partition $\mathcal{Z}$ into the product $\mathcal{Z} = \mathcal{Z}_0 \times \mathcal{X}$, where $\mathcal{Z}_0 := \{z_{01}, \ldots, z_{0K}\}$ and $\mathcal{X} := \{x_1, \ldots, x_L\}$. Now consider the following sequence of moment inequalities:

$$\mathbb{E}[(D - g(z_0, x))\, \mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.17}$$

$$\mathbb{E}[(g(z_0, x) - D)\, \mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.18}$$

and:

$$\mathbb{E}[(\mathbb{1}\{U \leq g(z_0, x)\} - g(z_0, x))\, \mathbb{1}\{X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.19}$$

$$\mathbb{E}[(g(z_0, x) - \mathbb{1}\{U \leq g(z_0, x)\})\, \mathbb{1}\{X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}. \tag{3.20}$$

Together (3.17) and (3.18) imply $P(D = 1|Z = z) = g(z)$ for all $z \in \mathcal{Z}$, and (3.19) and (3.20) imply $P(U \leq g(z)|Z = z) = P(U \leq g(z)|X = x) = g(z)$ for all $z_0 \in \mathcal{Z}_0$ and $x \in \mathcal{X}$. Under finiteness of the support $\mathcal{Z}$, these moment inequalities represent the only observable implications of the independence

condition $U \perp\!\!\!\perp Z_0|X$*. In addition, we will impose the following moment conditions:*

$$\mathbb{E}\left[t(z_0,x) - \mathbb{1}\{Z = z_0, X = x\}\right] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\ \forall x \in \mathcal{X}, \tag{3.21}$$

$$\mathbb{E}\left[\mathbb{1}\{Z = z_0, X = x\} - t(z_0,x)\right] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\ \forall x \in \mathcal{X}, \tag{3.22}$$

and:

$$\mathbb{E}\left[U_d\left(\mathbb{1}\{Z = z_0, X = x\}\sum_{z_0 \in \mathcal{Z}_0} t(z_0,x) - \mathbb{1}\{X = x\}t(z_0,x)\right)\right] \leq 0,\ \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X},\ d \in \{0,1\}, \tag{3.23}$$

$$\mathbb{E}\left[U_d\left(\mathbb{1}\{X = x\}t(z_0,x) - \mathbb{1}\{Z = z_0, X = x\}\sum_{z_0 \in \mathcal{Z}_0} t(z_0,x)\right)\right] \leq 0,\ \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X},\ d \in \{0,1\}. \tag{3.24}$$

Together (3.21) - (3.24) *imply the mean independence condition:* $\mathbb{E}[U_d|Z] = \mathbb{E}[U_d|X]$ *for* $d \in \{0,1\}$*. In particular,* (3.21) *and* (3.22) *ensure* $t(z_0,x) = P(Z_0 = z_0, X = x)$*, so that the moment conditions in* (3.23) *and* (3.24) *imply:*

$$\mathbb{E}\left[U_d\left(\mathbb{1}\{Z = z_0, X = x\}P(X = x) - \mathbb{1}\{X = x\}P(Z_0 = z_0, X = x)\right)\right] = 0,\ \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X},\ d \in \{0,1\},$$

or equivalently:

$$\mathbb{E}\left[U_d\left(\frac{\mathbb{1}\{Z = z_0, X = x\}}{P(Z_0 = z_0, X = x)} - \frac{\mathbb{1}\{X = x\}}{P(X = x)}\right)\right] = 0,\ \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X},\ d \in \{0,1\}.$$

From here, a full verification of Assumption 3.2.2 for the factual domain, including Effros measurability of the multifunction (3.16)*, is provided in Appendix 3.C.2.*

With this setup, we might be interested in how the outcome variable changes when the factors Z *that determine an individual's treatment decision are modified. For example, let* Γ *denote the set of all measurable functions* $\gamma : \mathcal{Z} \to \mathcal{Z}$ *(note that there are at most finitely many).*[15] *We can then define:*

$$Y_\gamma^\star = U_0(1 - D_\gamma^\star) + U_1 D_\gamma^\star, \tag{3.25}$$

where the random variable $D_\gamma^\star$ *is then given by:*

$$D_\gamma^\star = \mathbb{1}\{g_0(\gamma(Z)) \geq U\}.$$

Note that as in Heckman and Vytlacil (1999) and Heckman and Vytlacil (2005), our counterfactual $\gamma \in \Gamma$ *has no direct effect on* (U_0, U_1)*. Our interest is in the properties of the random variable* $Y_\gamma^\star$*, such as its mean or its conditional mean. The multifunction for the counterfactual domain is given by:*

$$\boldsymbol{G}^\star(Z, U_0, U_1, U, \theta, \gamma) := \left\{(Y_\gamma^\star, D_\gamma^\star) \in \mathcal{Y} \times \{0,1\} : \begin{array}{l} Y_\gamma^\star = U_0(1 - D_\gamma^\star) + U_1 D_\gamma^\star, \\ D_\gamma^\star = \mathbb{1}\{g(\gamma(Z)) \geq U\}. \end{array}\right\}. \tag{3.26}$$

Note here we take $\mathcal{Y}^\star = \mathcal{Y}$*. Again, close inspection of this multifunction provides some simplification:*

$$\boldsymbol{G}^\star(Z, U_0, U_1, U, \theta, \gamma) = \begin{cases} (U_1, 1), & \text{if } U \leq g(\gamma(Z)), \\ (U_0, 0), & \text{if } g(\gamma(Z)) < U. \end{cases} \tag{3.27}$$

A full verification of Assumption 3.2.3 for the counterfactual domain, including Effros measurability of the

[15] See Carneiro et al. (2011) for a discussion of other possible parameters under this setting.

multifunction (3.27)*, is provided in Appendix 3.C.2.*

3.2.3 The Policy Transform and Decision Problem

Throughout the paper we will build on the environment established in the previous section to present a framework for making policy decisions based on the value of any counterfactual object of interest that can be written as an integral of some function of the vector V_γ. In particular, if $\varphi : \Omega \times \Gamma \to \mathbb{R}$ is some measurable function, then we will restrict attention to environments where policymakers are interested in either the *policy transform* or the *conditional policy transform* of φ, which are defined next.

Definition 3.2.1 (Policy Transform and Conditional Policy Transform)**.** *Let $\varphi : \Omega \times \Gamma \to \mathbb{R}$ be a bounded and measurable function. The policy transform of φ is a function $I[\varphi](\gamma) : \Gamma \to \mathbb{R}$ given by:*

$$I[\varphi](\gamma) := \int \varphi(\omega, \gamma)\, dP. \tag{3.28}$$

Furthermore, if $\mathfrak{A}' \subset \mathfrak{A}$ is a $\sigma-$algebra, then a conditional policy transform of φ given $\mathfrak{A}'$ is a function $\tilde{I}[\varphi] : \Omega \times \Gamma \to \mathbb{R}$ such that (i) $\tilde{I}[\varphi] : \Omega \times \Gamma \to \mathbb{R}$ is $\mathfrak{A}' \otimes \Gamma-$measurable, and (ii) $I[\tilde{I}[\varphi](\cdot, \gamma)\mathbb{1}_A](\gamma) = I[\varphi \mathbb{1}_A](\gamma)$ for every $A \in \mathfrak{A}'$.

We will focus on the unconditional policy transform throughout the remainder of the paper, since analogous results hold for the conditional policy transform. In addition, since the relevant random variables in our environment are given in the vector V_γ, we will abuse notation throughout the paper and instead focus on policy transforms of the form:

$$I[\varphi](\gamma) := \int_\Omega \varphi(V_\gamma(\omega))\, dP = \int_{\mathcal{V}} \varphi(v)\, dP_{V_\gamma}, \tag{3.29}$$

which are clearly a special case of the general policy transforms in Definition 3.2.1.

In the remainder of the paper we take as primitive that the policymaker would like to choose γ to maximize the value of the policy transform for some known function $\varphi : \mathcal{V} \to \mathbb{R}$, although all results apply equally to the case where the policymaker wishes to minimize the policy transform.[16] For pedagogical purposes, it is useful to first consider an idealized decision problem. In particular, when (i) the true distribution $P_{Y,Z}$ is known, (ii) the conditional distribution $P_{U|Y,Z}$ is known, and (iii) the counterfactual conditional distribution $P_{Y^\star_\gamma|Y,Z,U}$ is known, the policymaker's problem becomes trivial: she can simply compute the policy transform of φ and choose the maximizing value of γ. However, clearly such idealized environments will be rare. Instead, we will consider the more realistic case when the policymaker only has access to an i.i.d. sample of size n from the true distribution $P_{Y,Z}$, and knows only that Assumptions 3.2.1, 3.2.2, and 3.2.3 are satisfied. In such an environment, the policymaker may be unable to compute the policy transform due to (i) lack of perfect knowledge of $P_{Y,Z}$, (ii) lack of knowledge of $P_{U|Y,Z}$ and (iii) lack of knowledge of $P_{Y^\star_\gamma|Y,Z,U}$. All three cases can occur when the structural parameters are point- or partially-identified.

We are now ready to define the decision problem under consideration.

We are now ready to define the decision problem under consideration.

Definition 3.2.2 (The Decision Problem)**.** *The policymaker's decision problem is characterized by:*

(i) The population, represented by the probability space $(\Omega, \mathfrak{A}, P)$.

(ii) The action (or policy) space, given by $(\Gamma, \mathfrak{B}(\Gamma))$.

[16] After we describe the decision problem, it will be apparent that desire of the policymaker to maximize or minimize the policy transform might be deduced using an axiomatic approach from a preference relation over the space of Borel probability measures on $\mathcal{V}$. We find this idea interesting, but will not pursue it here.

(iii) *The sample space, given by* $(\Psi_n, \Sigma_{\Psi_n}, P_{Y,Z}^{\otimes n})$*, where* $\Psi_n := (\mathcal{Y} \times \mathcal{Z})^n$*, with typical element* $\psi = \{(y_i, z_i)\}_{i=1}^n$*, equipped with the product Borel* $\sigma-$*algebra* $\Sigma_{\Psi_n} := (\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}))^{\otimes n}$ *and the product measure* $P_{Y,Z}^{\otimes n}$*.*

(iv) *The state space, given by* $\mathcal{S} \times \mathcal{P}_{Y,Z}$*, where* $\mathcal{P}_{Y,Z}$ *is the set of all Borel probability measures on* $\mathcal{Y} \times \mathcal{Z}$*, and* $\mathcal{S}$ *is the set of all triples* $s = (\theta, P_{U|Y,Z}, P_{Y_\gamma^\star|Y,Z,U})$ *such that the pair* $(s, P_{Y,Z})$ *satisfies:*

(a) $\theta \in \Theta$*,*

(b) $P_{U|Y,Z}(U \in \boldsymbol{G}^-(Y, Z, \theta) \,|Y = y, Z = z) = 1$*,* (y, z)*-a.s.,*

(c) $P_{Y_\gamma^\star|Y,Z,U}(Y_\gamma^\star \in \boldsymbol{G}^\star(Y, Z, U, \theta, \gamma)|Y = y, Z = z, U = u) = 1$*,* (y, z, u)*-a.s., and*

(d) the elements $\theta \in \Theta$ *and* $P_{U|Y,Z}$ *satisfy:*

$$\max_{j=1,\ldots,J} \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}}[m_j(Y, Z, U, \theta)] \leq 0. \tag{3.30}$$

(v) *The feasible statistical decision rules* $\mathcal{D}$*, with typical element* d*, given by the set of all measurable functions* $d : \Psi_n \to \Gamma$*.*

(vi) *The objective function, given by a function* $I[\varphi] : \Gamma \times \mathcal{S} \times \mathcal{P}_{Y,Z} \to \mathbb{R}$*, called the state-dependent policy transform, which has the expression:*

$$I[\varphi](\gamma, s) := \int \varphi(v)\, d(P_{Y_\gamma^\star|Y,Z,U} \times P_{U|Y,Z} \times P_{Y,Z}) \tag{3.31}$$

where $\varphi : \mathcal{V} \to \mathbb{R}$ *is a measurable function (where* $P_{Y,Z}$ *is left implicit when writing* $I[\varphi](\gamma, s)$*).*

A few remarks on this definition of our statistical decision problem are in order. In parts (i) and (ii), the specification of the population and the action space are somewhat standard, and have been motivated in the previous sections. In part (iii), the sample space is simply taken as the $n-$fold product of the observable space $(\mathcal{Y} \times \mathcal{Z})$. The measure on this space is the $n-$fold product of the true distribution $P_{Y,Z}$, from which we immediately deduce that the sample in $\psi \in \Psi_n$ is assumed to be i.i.d. Motivated from the framework in the previous section, part (iv) indicates that the unobserved state is characterized by a distribution $P_{Y,Z}$ and the triple $(\theta, P_{U|Y,Z}, P_{Y_\gamma^\star|Y,Z,U})$, where $\mathcal{S}$ corresponds to the set of all such triples that satisfy the model support restrictions and moment conditions introduced in the previous section. In part (v), the feasible decision rules $\mathcal{D}$ are characterized by the set of all measurable functions from the sample space Ψ_n to the action space Γ. We will return to this point below.[17] Furthermore, in this paper we will use the terms *policy rules* and *decision rules* interchangeably. Finally, part (vi) of Definition 3.2.2 introduces the state-dependent policy transform, which is a generalization of the policy transform that allows for it's value to depend on the unknown state from part (iv). Evaluated at the true state, the state-dependent policy transform reduces to the policy transform from Definition 3.2.1.

Ex-ante (i.e. before observing the sample) each decision rule $d : \Psi_n \to \Gamma$ is a random variable. Under some measurability conditions, this implies the state-dependent policy transform $I[\varphi](d(\psi), s)$ is also a random variable. The remaining question is how to use the collection $\{I[\varphi](d(\psi), s) : (s, P_{Y,Z}) \in \mathcal{S} \times \mathcal{P}_{Y,Z}\}$ to evaluate a given policy rule. It seems self-evident that a policy rule $d \in \mathcal{D}$ should be preferred to a policy rule $d' \in \mathcal{D}$ if for every $P_{Y,Z} \in \mathcal{P}_{Y,Z}$ we have $I[\varphi](d'(\psi), s) \leq I[\varphi](d(\psi), s)$ a.s. for every $s \in \mathcal{S}$; in such a case, d delivers a larger value of the policy transform in every state with probability one, regardless of the

[17] Note we might instead allow for randomized decision rules by taking $\mathcal{D}$ to be the set of all measurable functions from Ψ to the set of all *distributions* on Γ. This is not required for what we have in mind, but is easily accommodated under slightly modified assumptions.

distribution $P_{Y,Z}$. Any preference relation over $\mathcal{D}$ that satisfies this condition will be said to *respect weak dominance*.[18] However, beyond the requirement that a preference relation respect weak dominance, it is not obvious how a policymaker should (in the prescriptive sense) choose among competing policy options given the decision problem in Definition 3.2.2.[19]

Although a particular preference relation is not required in order to find the results in this paper interesting, it will be useful to define our notion of optimality in the policymaker's decision problem. In particular, our results may be especially useful to policymakers that are sympathetic to the following preference relation:

Definition 3.2.3 (PAC Maximin Preference Relation)**.** *Fix a sample size n. For any $\kappa \in (0,1)$ and any $d \in \mathcal{D}$, let $c_n(\cdot,\kappa) : \mathcal{D} \to \mathbb{R}_{++}$ be the smallest value satisfying:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) + c_n(d,\kappa) \geq \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) \right) \geq \kappa. \tag{3.32}$$

Then decision rule $d : \Psi_n \to \Gamma$ is weakly preferred to (or weakly dominates) decision rule $d' : \Psi_n \to \Gamma$ at level κ and sample size n, denoted by $d' \preccurlyeq_\kappa d$, if and only if $c_n(d,\kappa) \leq c_n(d',\kappa)$. The decision rule $d : \Psi_n \to \Gamma$ is strictly preferred to (or strictly dominates) decision rule $d' : \Psi_n \to \Gamma$, denoted by $d' \prec_\kappa d$, if and only if $c_n(d,\kappa) < c_n(d',\kappa)$. A decision rule $d \in \mathcal{D}$ will be called admissible with respect to $\preccurlyeq_\kappa$ if there is no decision rule $d' \in \mathcal{D}$ that is strictly preferred to (or strictly dominates) d.

This preference relation is named the PAC maximin preference relation given its close connection to the learning framework in the next subsection, which in turn is closely related to the PAC learning model of Valiant (1984) from computational learning theory. We refer readers to Appendix 3.A.2 where we discuss the notion of PAC learnability from computational learning theory. We will also emphasize the connection further in the next subsection.

For a fixed $\kappa \in (0,1)$, the preference relation from Definition 3.2.3 is a total ordering, meaning any two decision rules d and d' can be compared according to $\preccurlyeq_\kappa$. In addition, it has an interpretation in terms of quantiles. In particular, suppose for simplicity that $\mathcal{P}_{Y,Z}$ contains a single distribution π and define $Q_\pi(\kappa, d)$ as the κ quantile (under distribution π) of the map:

$$d \mapsto \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s). \tag{3.33}$$

Note that the map in (3.33) is always positive. Then a decision rule $d \in \mathcal{D}$ will be preferred to a decision rule $d' \in \mathcal{D}$ under $\preccurlyeq_\kappa$ if and only if $Q_\pi(\kappa, d) \leq Q_\pi(\kappa, d')$. Quantile utility maximization has been considered in Manski (1988) and Manski and Tetenov (2014), and axiomatized in Rostek (2010). However, our approach has major differences from these approaches, especially with regards to our treatment of the (sub-)states $s \in \mathcal{S}$.

Providing an axiomatization for the preference relation in Definition 3.2.3 is beyond the scope of this paper. Indeed, there is no reason why a policymaker needs to have the exact preference relation from Definition 3.2.3 in order to find the results in this paper useful or interesting. However, the following result shows that, at a minimum, $\preccurlyeq_\kappa$ respects weak dominance, as defined above.

Proposition 3.2.1. *Suppose that Assumptions 3.2.1, 3.2.2 and 3.2.3 hold, and that $\varphi : \mathcal{V} \to [\varphi_{lb}, \varphi_{ub}] \subseteq \mathbb{R}$ is a bounded and measurable function. Also, suppose that $\gamma \mapsto \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s)$ is (universally) measurable. Let*

[18] We refer to Manski (2011) for a similar definition. Also note that our definition implies stochastic dominance of $I[\varphi](d(\psi), s)$ over $I[\varphi](d'(\psi), s)$ for every $(s, P_{Y,Z}) \in \mathcal{S} \times \mathcal{P}_{Y,Z}$. By Strassen's Theorem, our definition will be equivalent to stochastic dominance if we allow for alternative probability spaces for each $(s, P_{Y,Z})$ pair.

[19] This point is raised repeatedly in the work of Charles Manski, and is summarized in Manski (2011).

$d, d' \in \mathcal{D}$ be two decision rules, and suppose that for every $P_{Y,Z} \in \mathcal{P}_{Y,Z}$ we have $I[\varphi](d'(\psi), s) \leq I[\varphi](d(\psi), s)$ a.s. for every $s \in \mathcal{S}$. Then for any $\kappa \in (0,1)$ we have $d' \preccurlyeq_\kappa d$, where $\preccurlyeq_\kappa$ is the preference relation from Definition 3.2.3; that is, the preference relation $\preccurlyeq_\kappa$ respects weak dominance.

Proof. See Appendix 3.B. ■

Remark 3.2.2. *Universal measurability is a weaker requirement then Borel measurability, and is defined in Appendix 3.B.2. Also, in Appendix 3.B.2 we show that the map $\gamma \mapsto \inf_{s\in\mathcal{S}} I[\varphi](\gamma, s)$ is universally measurable, although the result and proof relies on Assumption 3.3.1 introduced in the next section. Since Assumption 3.3.1 has not yet been introduced at this point, we impose (universal) measurability of $\gamma \mapsto \inf_{s\in\mathcal{S}} I[\varphi](\gamma, s)$ as a separate assumption in this proposition.*

Our main interest in the preference relation from Definition 3.2.3—especially versus other preference relations encountered in frequentist decision theory—is its close connection to the PAC learning framework, which allows us to use a rich set of results from statistical learning theory and empirical process theory to study its theoretical properties. Before formally introducing this connection, we will first revisit our examples to illustrate the various definitions presented in Definition 3.2.2.

Example 1 (Simultaneous Discrete Choice (Cont'd))**.** *For the simultaneous discrete choice example, recall that our interest is in the properties of the counterfactual random variable $Y^\star_{k,\gamma}$, such as its mean or its conditional mean. For the sake of illustration, we will focus on the quantity:*

$$I[\varphi](\gamma) = \int_\Omega \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}\, dP, \tag{3.34}$$

which is a counterfactual choice probability. Note this quantity is the policy transform of the function $\varphi(\omega, \gamma) = \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}$. Without much additional complication, we might instead be interested in the conditional choice probability $\mathbb{E}[\mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}|Z]$; it is easily verified that $\tilde{I}[\varphi](\omega, \gamma) = \mathbb{E}[\varphi(\omega, \gamma)|Z](\omega)$, with $\varphi(\omega, \gamma) = \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}$, is a conditional policy transform.[20] Throughout we will suppose the policymaker is interested in selecting the policy $\gamma \in \Gamma$ that maximizes the quantity (3.34). We can now formally define the policymaker's decision problem. The population is given by the probability space $(\Omega, \mathfrak{A}, P)$ and the action space is given by $(\Gamma, \mathfrak{B}(\Gamma))$, where Γ is the set of all functions $\gamma = (\gamma_k)_{k=1}^K$ with $\gamma_k : \mathcal{Z} \times \mathcal{Y}^{K-1} \to \mathcal{Z} \times \mathcal{Y}^{K-1}$ and $\mathfrak{B}(\Gamma)$ can be taken as the power set of Γ.[21] The sample space in this example is given by Ψ_n, which is all possible realizations of the n vectors $\{(y_i, z_i)\}_{i=1}^n$. Each state of the world is indexed by a pair $(\theta, P_{U|Y,Z})$ satisfying the support restriction given by (3.8) and the moment conditions (3.9) and (3.10). The state dependent policy transform is given by:

$$I[\varphi](\gamma, s) := \int \mathbb{1}\{U_k \leq \pi_k(\gamma(Z_k, Y_{-k}); \theta)\}\, dP_{U|Y,Z} dP_{Y,Z}.$$

A feasible statistical decision rule is then any measurable function $d : \Psi_n \to \Gamma$ that selects a policy indexed by γ given access to an $n-$sample from Ψ_n.

Example 2 (Program Evaluation (Cont'd))**.** *For the program evaluation example, recall that our interest is in the properties of the random variable $Y^\star_\gamma$, such as its mean or its conditional mean. For the sake of illustration, we will focus on the average outcome under some counterfactual policy $\gamma \in \Gamma$, given by*

[20] Indeed, by definition this quantity is measurable with respect to $\sigma(Z)$, and satisfies:

$$I[\tilde{I}[\varphi](\cdot, \gamma)\mathbb{1}_A](\gamma) = \int \mathbb{E}[\varphi(\omega, \gamma)|Z](\omega)\mathbb{1}_A(\omega)\, dP = \int \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}\mathbb{1}_A(\omega)\, dP = I[\varphi\mathbb{1}_A](\gamma), \tag{3.35}$$

for every $A \in \sigma(Z)$.

[21] Since $\mathcal{Z}$ and $\mathcal{Y}$ are finite, both Γ and $\mathfrak{B}(\Gamma)$ contain at most finitely many elements.

$\mathbb{E}[Y^\star_\gamma]$. *Note that taking* $\varphi(\omega,\gamma) = Y^\star_\gamma(\omega)\,(:= Y^\star(\omega,\gamma))$, *it is then clear that* $\mathbb{E}[Y^\star_\gamma] = I[\varphi](\gamma)$, *so that the average effect of a counterfactual policy is the policy transform of the random variable* $Y^\star_\gamma(\omega)$. *Without much additional complication, we might instead be interested in the conditional average effect* $\mathbb{E}[Y^\star_\gamma|X]$. *It is easily verified that* $\tilde{I}[\varphi](\omega,\gamma) = \mathbb{E}[\varphi(\omega,\gamma)|X](\omega)$, *with* $\varphi(\omega,\gamma) = Y^\star_\gamma(\omega)$, *is a conditional policy transform.*[22] *We will assume throughout that the policymaker is interested in maximizing the value of* $\mathbb{E}[Y^\star_\gamma]$. *We can now formally define the policymaker's decision problem. The population is given by the probability space* $(\Omega, \mathfrak{A}, P)$ *and the action space is given by* $(\Gamma, \mathfrak{B}(\Gamma))$, *where* Γ *is the set of all functions* $\gamma : \mathcal{Z} \to \mathcal{Z}$ *and* $\mathfrak{B}(\Gamma)$ *is the power set of* Γ.[23] *The sample space is given by* $\Psi_n = (\mathcal{Y} \times \{0,1\} \times \mathcal{Z})^n$ *with a typical element* $\psi = ((y_i, d_i, z_i))_{i=1}^n$. *The state space* $\mathcal{S}$ *is given by* $s = (\theta, P_{U_0,U_1,U|Y,Z}, P_{Y^\star_\gamma|U_0,U_1,U,Y,Z})$, *where* $P_{U_0,U_1,U|Y,Z}$ *and* $P_{Y^\star_\gamma|U_0,U_1,U,Y,Z}$ *are any random variables that satisfy the support restriction* (3.15) *and moment conditions* (3.17) - (3.22). *Finally, a feasible statistical decision rule is any measurable function* $d : \Psi_n \to \Gamma$ *that selects a policy indexed by* γ *given access to an* $n-$*sample from* Ψ_n.

3.2.4 A Roadmap to the Theoretical Results: Ex-ante and Ex-post Analyses

With the policymaker's decision problem defined in the previous subsection, our upcoming theoretical results can be divided according to whether they are applicable ex-ante (i.e. before observing the sample) or ex-post (i.e. after observing the sample).

Recall the preference relation from Definition 3.2.3. Under this preference relation, the "performance" or "quality" of a decision rule d can be measured using the value $c_n(d,\kappa)$. Thus, the value of $c_n(d,\kappa)$ will be a major focus of both the ex-ante and ex-post theoretical analyses in the remainder of the paper. Our main focus in the ex-ante theoretical results is establishing sufficient conditions for learnability of a policy space, which we will discuss further in this subsection. Our main focus for the ex-post theoretical analysis is in establishing bounds on the value of $c_n(d,\kappa)$ for certain decision rules, as well as bounds on the set of decision rules $d \in \mathcal{D}$ that obtain a small value of $c_n(d,\kappa)$.

Policy Space Learnability

To understand the ex-ante theoretical analysis, we must formally introduce the concept of policy space learnability, named because of its connection to notions of learnability from computational learning theory. Intuitively, a policy space Γ will be learnable if, for some decision rule $d \in \mathcal{D}$, the value $c_n(d,\kappa)$ from Definition 3.2.3 can be made arbitrarily small as n increases. This concept will be made precise in this subsection.

A review of concepts of learnability from computational learning theory is provided in Appendix 3.A.2. We argue that, under the preference relation from Definition 3.2.3, the conceptual differences between the problem of policy choice and the problem of selecting an optimal classifier in a statistical learning setting are smaller than they may initially appear. In both settings we wish to select a decision rule based on a finite sample that will perform well, based on similar criteria, in samples yet unseen. The essential difference between the environments is that the performance of a counterfactual policy is unobservable, even for the sample in hand. Of course this is not an issue if the policymaker has an econometric model that can used to determine the counterfactual outcomes of the policy experiment.

[22] Indeed, by definition this quantity is measurable with respect to $\sigma(X)$, and satisfies:

$$I[\tilde{I}[\varphi](\cdot,\gamma)\mathbb{1}_A](\gamma) = \int \mathbb{E}[\varphi(\omega,\gamma)|X](\omega)\mathbb{1}_A(\omega)\,dP = \int Y^\star_\gamma(\omega)\mathbb{1}_A(\omega)\,dP = I[\varphi\mathbb{1}_A](\gamma), \tag{3.36}$$

for every $A \in \sigma(X)$.

[23] Since $\mathcal{Z}$ is finite, both Γ and $\mathfrak{B}(\Gamma)$ contain at most finitely many elements.

The general model from the previous subsections will serve exactly this purpose. Given the preference relation from Definition 3.2.3, the policymaker is presented with a decision problem that is remarkably similar to a learning problem, which is apparent when the following definition is compared with the definition of PAC Learnability from Appendix 3.A.2.

Definition 3.2.4 (PAMPAC Learnability)**.** *Under Assumptions 3.2.1, 3.2.2, and 3.2.3, a policy space Γ is policy agnostic maximin PAC-learnable (PAMPAC) with respect to the policy transform of $\varphi : \mathcal{V} \to \mathbb{R}$ if there exists a function $\zeta_\Gamma : \mathbb{R}_{++} \times (0,1) \to \mathbb{N}$ such that, for any $(c, \kappa) \in \mathbb{R}_{++} \times (0,1)$ and any distribution $P_{Y,Z}$ over $\mathcal{Y} \times \mathcal{Z}$, if $n \geq \zeta_\Gamma(c, \kappa)$ then there is some decision procedure $d : \Psi_n \to \Gamma$ satisfying:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) + c \geq \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) \right) \geq \kappa. \tag{3.37}$$

That is, a policy space is PAMPAC learnable if there is exists some decision rule $d : \Psi_n \to \mathbb{R}$ that, in the worst-case (sub-)state $s \in \mathcal{S}$, closely approximates the value:

$$\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s),$$

with high probability for a sufficiently large (but finite) sample.[24] In terms of the preference relation from Definition 3.2.3, PAMPAC learnability implies that, as the sample size grows, every point in $(c, \kappa)-$space must eventually (i.e. for large enough n) lie above the function $c_n(d, \cdot) : (0,1) \to \mathbb{R}_{++}$ for some decision rule d. This idea is illustrated in Figure 3.3. Framed in this manner, we see that PAMPAC learnability is not required to determine the admissible decision rules or to make a policy choice. However, there may be substantial ex-ante limitations on the theoretical performance of any given decision rule in environments that are not PAMPAC learnable, making it an important object of theoretical analysis.

Despite the fact that PAMPAC learnability may appear to be a weak notion, there are trivial environments where a policy space Γ may not be PAMPAC learnable.

Example 1 (Simultaneous Discrete Choice (cont'd))**.** *Consider the general setup of Example 1. Suppose for simplicity that $K = 1$, and consider the following modifications. Let $\mathcal{Z} = [-1, 1]$ and $\Theta = [-1, 1]$ and let $\pi_k(Z_k, Y_{-k}; \theta) = \pi_k(Z_k; \theta) = \sin(Z_k/\theta)$. Then Y_k is determined by the equation:*

$$Y_k = \mathbb{1}\{\sin(Z_k/\theta) \geq U_k\}.$$

Now consider a policy space Γ that consists of all functions $\gamma : \mathcal{Z} \to \mathcal{Z}$, and suppose we are interested in the policy transform:

$$I[\varphi](\gamma) := \int_\Omega \varphi(\omega, \gamma)\, dP = \int_\Omega \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}\, dP,$$

where $\varphi(\omega, \gamma) = \mathbb{1}\{Y^\star_{k,\gamma}(\omega) = 1\}$ and:

$$Y^\star_{k,\gamma} = \mathbb{1}\{\sin(\gamma(Z_k)/\theta) \geq U_k\}.$$

In this case, we claim the policy space Γ may not be PAMPAC learnable with respect to the policy transform

[24] A nearly identical definition can be given for policy agnostic minimax PAC-learnability, with the exception that the decision procedure $d : \Psi_n \to \Gamma$ must satisfy:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\sup_{s \in \mathcal{S}} I[\varphi](d(\psi), s) - c \leq \inf_{\gamma \in \Gamma} \sup_{s \in \mathcal{S}} I[\varphi](\gamma, s) \right) \geq \kappa. \tag{3.38}$$

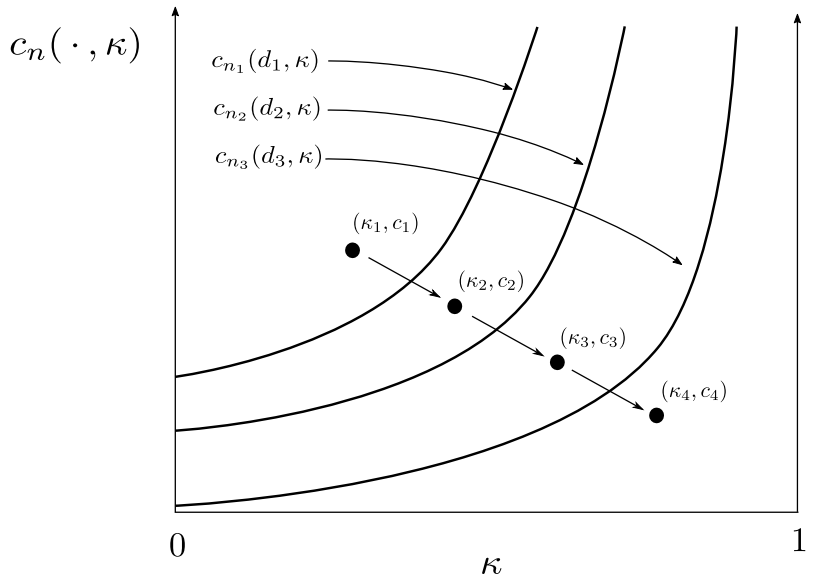


Figure 3.3: This figure illustrates the idea of PAMPAC learnability from Definition 3.2.4. Given a pair (c,κ), PAMPAC learnability guarantees that there is some finite n and some decision rule $d:\Psi_n\to\Gamma$ such that the graph of $c_n(d,\kappa)$ lies entirely below the point (c,κ). For example, for (c_1,κ_1) in the figure, there exists a sample size n_1 and a decision rule d_1 such that (3.37) is satisfied. Note that (3.37) is also satisfied for the points (c_2,κ_2) and (c_3,κ_3) at n_2 and d_2, and n_3 and d_3, respectively. To verify PAMPAC learnability, the same must hold for all points (c,κ); in particular, in the figure we would need to find a sample size n_4 and decision rule d_4 such that the graph of $c_{n_4}(d_4,\kappa)$ lies entirely below the point (c_4,κ_4).

of φ.

It is important to realize that the possible failure of PAMPAC learnability does not hinge on the choice of the sine function in this example, which is used for illustrative purposes only. Indeed, the following example shows that the idea is more general.

Example 2 (Program Evaluation (cont'd))**.** *Consider the general setup of Example 2, with the following modifications. Let $\mathcal{Z}=[-1,1]$ and let Θ denote the space of continuous functions with values in $[-1,1]$. Otherwise, keep all other aspects of the factual domain the same. Now consider a policy space Γ that consists of all continuous functions $\gamma:\mathcal{Z}\to\mathcal{Z}$. Suppose we are still interested in the policy transform of $\varphi(\omega,\gamma)=Y^\star_\gamma(\omega)$, where:*

$$Y^\star_\gamma = U_0(1-D^\star_\gamma)+U_1D^\star_\gamma, \tag{3.39}$$

and where the random variable $D^\star_\gamma$ is given by:

$$D^\star_\gamma = \mathbb{1}\{\theta_0(\gamma(Z))\geq U\}.$$

In this case, we claim the policy space Γ may not be PAMPAC learnable with respect to the policy transform of φ.

These examples illustrate that there may be limits to which policy spaces are learnable. In the first example, learnability may fail because the structural function determining the counterfactual values of $Y^\star_{k,\gamma}$ is too "complex," and so cannot be adequately approximated (or "learned") with any finite amount of data. A similar explanation applies to the second example, in particular to the structural function determining the values of $D^\star_\gamma$. In the next sections we will explore sufficient conditions for the learnability of a policy space that are precisely related to constraints on the complexity of certain function spaces. After establishing a

particular policy space is learnable, which is an ex-ante (i.e. before observing the sample) notion, we will then discuss how to evaluate particular decision rules, which is an ex-post (i.e. after observing the sample) notion. Both components will be relevant to the theoretical evaluation of the decision problem.

3.2.5 The Path Forward

As is suggested by (3.37) in Definition 3.2.4, and as was discussed in the introduction, in order to determine whether a given policy space Γ is PAMPAC learnable it is useful to first provide a characterization of the envelope functions:

$$I_{\ell b}[\varphi](\gamma) := \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s), \qquad I_{ub}[\varphi](\gamma) := \sup_{s \in \mathcal{S}} I[\varphi](\gamma, s).$$

Note that, at the true distribution $P_{Y,Z}$, the function $I_{\ell b}[\varphi](\gamma)$ serves as a lower bound on the policy transform $I[\varphi](\gamma)$. Similarly, the function $I_{ub}[\varphi](\gamma)$ serves as an upper bound. Recall that this idea was illustrated in Figure 3.1 in the introduction.

In the case of PAMPAC learnability, if a tractable characterization of the lower envelope function $I_{\ell b}[\varphi](\gamma)$ can be provided under some conditions, then determining whether a policy space is PAMPAC learnable reduces to the problem of finding a decision rule $d : \Psi_n \to \Gamma$ that satisfies:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\sup_{\gamma \in \Gamma} I_{\ell b}[\varphi](\gamma) - I_{\ell b}[\varphi](d(\psi)) \leq c \right) \geq \kappa, \tag{3.40}$$

for large enough (but finite) n. Thus in the next section we focus on obtaining a tractable characterization of the envelope functions before returning to the problem of policy choice in Section 3.4. Once a tractable characterization of the lower (or upper) envelope function is provided, we will then present sufficient conditions for PAMPAC learnability. In addition to its importance to our ex-ante analysis, we will see that a tractable characterization of the envelope functions will also be key to our ex-post analysis of the policymaker's decision problem in Section 3.5.

3.3 Envelope Functions for the Policy Transform

3.3.1 Preliminaries

In this section we derive a useful characterization of the envelope functions $I_{\ell b}[\varphi](\gamma)$ and $I_{ub}[\varphi](\gamma)$ defined in the previous section. We will show that these envelope functions can be written as the value functions of optimization problems parameterized by $\gamma \in \Gamma$. Our specific characterization will be important when deriving our learnability results, as well as for our ex-post finite-sample analysis in the next sections. However, for those interested in partial identification, the results in this section may be of substantial separate interest.

We first define the identified set for the structural parameters and policy transform before presenting our main result for this section. In general, these identified sets must be defined *relative* to a distribution $P_{Y,Z}$.[25] For notational simplicity this is kept implicit throughout this section.

We now begin by introducing some additional notation. For assistance with some of the notation in the next definition, the reader is referred to Appendix 3.A, which discusses the notion of selectionability from a random set.

Definition 3.3.1 (Distributions of Selections)**.** *The collection $\mathcal{P}_{U|Y,Z}(\theta)$ contains all regular conditional probability measures $P_{U|Y,Z}$ such that each $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ is the distribution of some selection $U \in$*

[25] See, for example, Definition 3 in Chesher and Rosen (2017a) and the surrounding discussion.

$Sel(\boldsymbol{G}^-(\cdot,\theta))$; that is:[26]

$$\mathcal{P}_{U|Y,Z}(\theta) := \left\{P_{U|Y,Z} : U \sim P_{U|Y,Z} \text{ for some } U \in Sel(\boldsymbol{G}^-(\cdot,\theta))\right\}. \tag{3.41}$$

Furthermore, the collection $\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)$ contains all regular conditional probability measures $P_{Y^\star_\gamma|Y,Z,U}$ such that each $P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)$ is the distribution of some selection $Y^\star_\gamma \in Sel(\boldsymbol{G}^\star(\cdot,\theta,\gamma))$; that is:[27]

$$\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma) := \left\{P_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma) : Y^\star_\gamma \sim P_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma) \text{ for some } Y^\star_\gamma \in Sel(\boldsymbol{G}^\star(\cdot,\theta,\gamma))\right\}. \tag{3.42}$$

We will see shortly that compactness of $\mathcal{U}$ from Assumption 3.2.1 is quite convenient. Indeed, note that under compactness of $\mathcal{U}$, the collection $\mathcal{P}_{U|Y,Z}(\theta)$ is uniformly tight for any θ. If $\mathcal{P}_{U|Y,Z}(\theta)$ is also closed in the weak* topology, then the collection $\mathcal{P}_{U|Y,Z}(\theta)$ is compact in the weak* topology, which allows for a simplification of the statement and proofs of many of the results. However, by the fact that $\boldsymbol{G}^-$ is closed, this latter result follows directly from the fact that every selection $U \in Sel\left(\boldsymbol{G}^-(\cdot,\theta)\right)$ is supported by a compact set.[28] Thus, throughout our exposition we can use the fact that $\mathcal{P}_{U|Y,Z}(\theta)$ is compact in the weak* topology.

Beyond the simplifications that come with this result, it also solves a meaningful issue related to selections from identically distributed random sets. Indeed, two identically distributed random sets may have different sets of measurable selections, although the weak* closure of their measurable selections will always coincide.[29] The issue is thus entirely resolved by compactness of $\mathcal{U}$, which ensures the collection $\mathcal{P}_{U|Y,Z}(\theta)$ is closed in the weak* topology; in other words, under Assumptions 3.2.1 and 3.2.2, this means two identically distributed random sets $\boldsymbol{G}^-(Y,Z,\theta)$ and $\boldsymbol{G}^-(Y',Z',\theta)$ (see Definition 3.A.2 in Appendix 3.A) will have the same set of measurable selections.

With the additional notation afforded by Definition 3.3.1, we now have the following definition of the identified set of structural parameters:

Definition 3.3.2 (Identified Set of Structural Parameters)**.** *Under Assumptions 3.2.1 and 3.2.2, the identified set Θ^* of structural parameters (with respect to the distribution $P_{Y,Z}$) is given by:*

$$\Theta^* := \left\{\theta \in \Theta : \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}}\left[m_j(y,z,u,\theta)\right] \leq 0\right\}. \tag{3.43}$$

Compactness of $\mathcal{P}_{U|Y,Z}(\theta)$ in the weak* topology, combined with boundedness of the moment conditions, ensures that the infimum in the definition of Θ^* is obtained.[30] Although our focus in this paper is not on the identified set of structural parameters, this definition will be helpful when providing a definition of the identified set for the policy transform, as well as in the proofs.

To state the definition of the identified set for the policy transform, it will be useful for us to first define the following function:

$$I^*[\varphi](\theta,\gamma,I,P_{Y^\star_\gamma|Y,Z,U},P_{U|Y,Z})$$

[26] Clearly the collection $\mathcal{P}_{U|Y,Z}(\theta)$ also depends on $P_{Y,Z}$, although we suppress this dependence for notational simplicity throughout.

[27] Clearly the collection $\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)$ also depends on $P_{Y,Z,U}$, although we suppress this dependence for notational simplicity throughout.

[28] See Corbae et al. (2009) Theorem 9.9.2 on p. 575, as well as the surrounding discussion.

[29] See Molchanov (2017) Theorem 1.4.3 on p. 79.

[30] This follows from the extreme value theorem after noting the map $P_{U|Y,Z} \mapsto \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}}\left[m_j(y,z,u,\theta)\right]$ is continuous when the moment function m_j is uniformly bounded.

$$:= \max\left\{ \left| \mathbb{E}_{P_{Y^\star_\gamma|Y,Z,U} \times P_{U|Y,Z} \times P_{Y,Z}} \left[\varphi(V_\gamma) - I \right] \right|, \max_{j=1,\ldots,J} \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}} \left[m_j(y,z,u,\theta) \right] \right\}. \tag{3.44}$$

Intuitively, this function is less than zero if and only if (i) all moment conditions are satisfied at the distribution $P_{Y,Z}$ and the pair $(\theta, P_{U|Y,Z})$, and (ii) if the point "I" is the resulting value of the policy transform for the inputs $(\theta, \gamma, P_{Y^\star_\gamma|Y,Z,U}, P_{U|Y,Z})$. As such, it represents all the conditions necessary for the point "I" to be included in the identified set for the policy transform. We now have the following definition:

Definition 3.3.3 (Identified Set for Policy Transforms)**.** *Under Assumptions 3.2.1, 3.2.2, and 3.2.3, for any $\gamma \in \Gamma$ the identified set for $I[\varphi](\gamma)$ (with respect to the distribution $P_{Y,Z}$) is given by:*

$$\mathcal{I}^*[\varphi](\gamma) := \bigcup_{\theta \in \Theta^*} \mathcal{I}[\varphi](\theta, \gamma), \tag{3.45}$$

where:

$$\mathcal{I}[\varphi](\theta,\gamma) := \Big\{ I \in \mathbb{R} : \exists P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta) \text{ and } P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma) \\ \text{satisfying } I^*[\varphi]\left(\theta, \gamma, I, P_{Y^\star_\gamma|Y,Z,U}, P_{U|Y,Z}\right) \leq 0 \Big\}. \tag{3.46}$$

Our main result in this section will attempt to provide a more insightful characterization of the identified set for policy transforms, which will also be vital for the problem of policy choice considered in the next section. However, before stating our main identification result, we require the following technical assumption.

Assumption 3.3.1 (Error Bounds)**.** *(i) (Linear Minorant) There exists values $\delta > 0$ and $C_1 > 0$ such that for every $\theta \in \Theta$:*

$$\inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}} \left[m_j(y,z,u,\theta) \right]|_+ \geq C_1 \min\{\delta, d(\theta, \Theta^*)\}. \tag{3.47}$$

(ii) (Local Counterfactual Robustness) There exists a value $C_2 \geq 0$ such that for any $\theta \in \Theta^*_\delta := \{\theta : d(\theta, \Theta^*) \leq \delta\}$:

$$\inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} \\ \geq \inf_{\theta^* \in \Theta^*} \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta^*)} \inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta^*,\gamma)} \int \varphi(v)\, dP_{V_\gamma} - C_2 d(\theta, \Theta^*), \tag{3.48}$$

and:

$$\sup_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \sup_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} \\ \leq \sup_{\theta^* \in \Theta^*} \sup_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta^*)} \sup_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta^*,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + C_2 d(\theta, \Theta^*). \tag{3.49}$$

Intuitively, Assumption 3.3.1 makes two statements. First, part (i) of the assumption is a global condition that requires that, whenever $\theta \in \Theta \setminus \Theta^*$, there is at least one moment function that can be bounded below by the function on the right side of (3.47). In general this condition is very similar to previous conditions in the literature; see, for example, the "partial identification condition" in Chernozhukov et al. (2007b) section 4.2. Also, see Kaido et al. (2019b) for a review of similar conditions. The major difference arises from the fact that the condition must hold for all $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$, owing to the fact that the moment conditions in this paper are allowed to depend on the latent variables. Verifying condition (i) can usually be done by

first enumerating all scenarios which imply $\theta \notin \Theta^*$, and then verifying that the condition holds for each such scenario. This is exactly the strategy used when verifying the assumption in the examples. Also note that the condition is automatically satisfied if $\mathcal{P}_{U|Y,Z}(\theta)$ is empty—that is, when $\boldsymbol{G}^-(Y,Z,\theta)$ admits no measurable selections—or when none of the moment conditions depend on the structural parameters.

Part (ii) of Assumption 3.3.1 appears to be entirely new. Intuitively, (3.48) is a local condition that requires the smallest value of the integral of φ to not decrease too fast as we move θ slightly outside of the identified set. In the opposite direction, (3.49) requires that the largest value of the integral of φ does not increase too fast as we move θ slightly outside of the the identified set. These conditions will be violated if, for example, the value of the integral can change discontinuously on the boundary of the identified set. We call the condition the *local counterfactual robustness* condition because it demands that small changes in the value of the structural parameters do not generate discontinuous changes in value of the counterfactual quantity of interest. Interestingly, both of the conditions in Assumption 3.3.1 are related to typical assumptions made in the theory of error bounds in the optimization literature.[31] Finally, note the value of δ in parts (i) and (ii) are the same. However, this is not restrictive, since part (i) and (ii) can be established for two different values $\delta_{(i)}, \delta_{(ii)} > 0$, and then δ can be taken as $\delta = \min\{\delta_{(i)}, \delta_{(ii)}\}$.

In practice, part (ii) of Assumption 3.3.1 can be challenging to verify. Because of this, we introduce the following assumption as an alternative to part (ii) of Assumption 3.3.1:

Assumption 3.3.2 (Error Bounds (2)(ii))**.** *For some $\delta > 0$, there exists values $\ell_1, \ell_2 \geq 0$ (possibly depending on δ) such that:*

$$d(u, \boldsymbol{G}^-(y,z,\theta)) \leq \ell_1 \cdot d(\theta, \Theta^-(y,z,u) \cap \Theta^*_\delta), \qquad (y,z) - a.s. \text{ for all } u \in \mathcal{U} \text{ and } \theta \in \Theta^*_\delta, \tag{3.50}$$

$$d(y^\star, \boldsymbol{G}^\star(y,z,u,\theta,\gamma)) \leq \ell_2 \cdot d(\theta, \Theta^\star(v,\gamma) \cap \Theta^*_\delta), \qquad (y,z,u) - a.s. \text{ for all } y^\star \in \mathcal{Y}^\star \text{ and } \theta \in \Theta^*_\delta. \tag{3.51}$$

where $\Theta^-(y,z,u)$ and $\Theta^\star(v,\gamma)$ are defined by:

$$\Theta^-(y,z,u) := \left\{\theta : u \in \boldsymbol{G}^-(y,z,\theta)\right\}, \qquad \Theta^\star(v,\gamma) := \left\{\theta : y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)\right\}.$$

Furthermore, the function $\varphi : \mathcal{V} \to \mathbb{R}$ is bounded, measurable, and Lipschitz continuous in $(u, y^\star)$ with Lipschitz constant L_φ.

The following Lemma shows that Assumption 3.3.2 is sufficient for part (ii) of Assumption 3.3.1. In the process, the Lemma makes an interesting connection between Assumption 3.3.1 and certain Lipschitzian behaviour of the random sets $\boldsymbol{G}^-$ and $\boldsymbol{G}^\star$ with respect to the structural parameters $\theta \in \Theta$.

Lemma 3.3.1. *Suppose that Assumptions 3.2.1, 3.2.2 and 3.2.3 are satisfied. Finally, suppose that $\boldsymbol{G}^-(\cdot,\theta)$ and $\boldsymbol{G}^\star(\cdot,\theta,\gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. Then Assumption 3.3.2 implies Assumption 3.3.1(ii) with $C_2 = L_\varphi \max\{\ell_1, \ell_2\}$.*

Proof. See Appendix 3.B. ■

It can be shown that the conditions (3.50) and (3.51) are equivalent to almost-sure versions of Lipschitz continuity conditions for set-valued maps, where the distance between two sets is measured by the Pompeiu–Hausdorff distance. Localized versions of these conditions are called *metric regularity* conditions, which also have a close connection to constraint qualifications from optimization theory. See Dontchev and Rockafellar (2009) Chapter 3.3 and Ioffe (2016) for a discussion.

[31] See Pang (1997) for an introduction.

3.3.2 Envelope Functions for the Policy Transform

We can finally turn to our main objective for this section, which is the problem of bounding the policy transform $I[\varphi](\gamma)$. Theoretically, bounds on $I[\varphi](\gamma)$ can be obtained by solving two (very) complicated constrained optimization problems that search over all distributions $P_{U|Y,Z}$ and $P_{Y^\star_\gamma|Y,Z,U}$ that satisfy our modelling assumptions for the ones that maximize and minimize the policy transform of φ. However, it is clear that such optimization problems will be infeasible in most realistic cases. The following result shows a tractable formulation of bounds on policy transforms that will be important for the next section.

Theorem 3.3.1 (Bounds on the Policy Transform)**.** *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3 and 3.3.1 all hold. Also, suppose that $\varphi : \mathcal{V} \to [\varphi_{\ell b}, \varphi_{ub}] \subset \mathbb{R}$ is a bounded, measurable function, and that for each $\gamma \in \Gamma$, the random sets $\boldsymbol{G}^-(\cdot,\theta)$ and $\boldsymbol{G}^\star(\cdot,\theta,\gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. Then $\overline{co}\,\mathcal{I}^*[\varphi](\gamma) = [I_{\ell b}[\varphi](\gamma), I_{ub}[\varphi](\gamma)]$, with:*

$$I_{\ell b}[\varphi](\gamma) = \inf_{\theta\in\Theta} \max_{\lambda_j\in\{0,1\}} \int \inf_{u\in\boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y,z,u,\theta)\right) dP_{Y,Z}, \tag{3.52}$$

$$I_{ub}[\varphi](\gamma) = \sup_{\theta\in\Theta} \min_{\lambda_j\in\{0,1\}} \int \sup_{u\in\boldsymbol{G}^-(y,z,\theta)} \sup_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) - \mu^* \sum_{j=1}^{J} \lambda_j m_j(y,z,u,\theta)\right) dP_{Y,Z}, \tag{3.53}$$

where $\mu^* \in \mathbb{R}_+$ is any value satisfying:

$$\mu^* \geq \max\left\{\frac{C_2}{C_1}, \frac{(\varphi_{ub} - \varphi_{\ell b})}{C_1\delta}\right\}, \tag{3.54}$$

and where C_1, C_2 and δ are from Assumption 3.3.1.

Proof. See Appendix 3.B. ■

Theorem 3.3.1 states that the closed, convex hull of the identified set $\mathcal{I}^\star[\varphi](\gamma)$ from Definition 3.3.3 for the policy transform $I[\varphi](\gamma)$ can be computed as the solution to two optimization problems. Interestingly, these optimization problems are closely related to problems found in the literature on mathematical programming problems subject to equilibrium constraints (MPECs), which have previously seen applications in economics to social planning problems and Stackelberg games.[32] The upper and lower envelope functions in Theorem 3.3.1 are perhaps most aptly characterized as penalized optimization problems, with μ^* in (3.54) serving the role of the penalty parameter. Both the statement of the result and its proof rely on the theory of exact penalty functions from the literature on error bounds in variational analysis.[33] The Theorem uses the error bounds Assumption 3.3.1 in order to show that the penalty μ^* can be taken to be finite. This is very important for the theoretical analysis of the policy decision problem to take place in the sections ahead. Furthermore, implicitly Theorem 3.3.1 shows that the values of λ_j will depend only on the parameter θ, a point which will used in the next sections.

From an identification perspective, the envelope functions will generally not give sharp bounds on the policy transform. However, under any additional conditions that ensure the identified set $\mathcal{I}^*(\gamma)$ is closed and convex for every $\gamma \in \Gamma$, Theorem 3.3.1 provides a (point-wise in γ) sharp characterization of the identified set for the policy transform. Finally, the result is easily modified for the case when the object of interest is a conditional policy transform.

One of the most interesting features of Theorem 3.3.1 is that, when the counterfactual object of interest is a particular form, there is no need to compute the identified set Θ^* of structural parameters in order to

[32] For a textbook treatment, see Luo et al. (1996).
[33] See Dolgopolik (2016) for a review.

bound the counterfactual object of interest. In addition, the unobservables in the problem are profiled out, and when the identified set $\mathcal{I}^*(\gamma)$ is closed and convex this is without any loss of information. This point also translates into the policy decision problem studied in the next sections. The structural parameters and unobservables intuitively play the role of an intermediary connecting the factual and counterfactual domains. However, after the envelope functions from Theorem 3.3.1 are computed, they play no further role in the problem of policy choice.

While we will not dwell on measurability issues in the main text, we note that Lemma 3.B.1 in Appendix 3.B.2 shows that the integrands in the optimization problems are universally measurable; that is, measurable for the completion of any probability measure $P_{Y,Z}$. The proof of this result relies crucially on the fact that both $\boldsymbol{G}^-$ and $\boldsymbol{G}^\star$ are Effros-measurable. Furthermore, Proposition 3.B.1 in Appendix 3.B.2 shows that the maps $\gamma \mapsto I_{\ell b}[\varphi](\gamma), I_{ub}[\varphi](\gamma)$ are measurable with respect to the universal $\sigma-$algebra on Γ (as generated by the Borel $\sigma-$algebra). These results will be important to keep in mind in the next sections on policy choice.

We now return to the examples presented earlier to discuss our identification result. We will first verify Assumption 3.3.1 in our examples and will show how Lemma 3.3.1 can be helpful.

Example 1 (Simultaneous Discrete Choice (cont'd)). *Consider again Example 1 on simultaneous discrete choice, and recall that we have imposed a median zero and median independence restriction using the moment conditions in (3.9) and (3.10).*

This example presents challenges for the verification of Assumption 3.3.1 because of the discontinuity of the function $\varphi(v) = \mathbb{1}\{\pi_k(\gamma(z, y_{-k}); \theta) \geq u\}$. Indeed, under our current assumptions, Assumption 3.3.1 is not satisfied. To appreciate the intuition, focus on Assumption 3.3.1(ii). The issue for this assumption arises only when for some $k \in \{1, \ldots, K\}$ and some $z \in \mathcal{Z}$ and $y_{-k} \in \mathcal{Y}^{K-1}$ we have (i) the counterfactual cutoff value $\pi_k(\gamma(z, y_{-k}); \theta^*) = 0$ at some $\theta^* \in \partial\Theta^*$, and if (ii) $P(Y_k = 1 | Z_k = z', Y_{-k} = y'_{-k}) \neq 0.5$, where $(z', y'_{-k}) = \gamma(z, y_{-k})$. In this knife-edge case, a very small change from $\theta^* \in \partial\Theta^*$ to some $\theta \notin \Theta^*$ can cause a discontinuous change in $P(Y^\star_{\gamma,k} = 1)$. A full description of this failure, including illustrations of various cases, is presented in Appendix 3.C.1.

However, by slightly strengthening our moment conditions we can satisfy Assumption 3.3.1 in this example. The key is to introduce additional assumptions on the degree of smoothness of the distribution of U_k around zero. In particular, we will replace the moment conditions in (3.9) and (3.10) with the following conditions:

$$\mathbb{E}\left[\left(\mathbb{1}\{U_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\} - 0.5\right) \mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}\right] \leq 0, \tag{3.55}$$

$$\mathbb{E}\left[\left(0.5 - \mathbb{1}\{U_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{-L_0 \pi_k(z', y'_{-k}; \theta), 0\}\right) \mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}\right] \leq 0, \tag{3.56}$$

for $k = 1, \ldots, K$, for all $z, z' \in \mathcal{Z}$ and all $y_{-k}, y'_{-k} \in \mathcal{Y}^{K-1}$. In addition to implying the median zero/median independence assumption, these new moment conditions also limit the amount of probability mass on $\mathcal{U}$ that is arbitrarily close to zero, which turns out to be key to satisfying Assumption 3.3.1. Also note that, despite the fact that these moment conditions will implicitly impose constraints on the obtainable counterfactual choice probabilities, it is easily verified that they do not impose any additional constraints on the set of structural parameters $\theta \in \Theta$ that can rationalize the observed distribution (in the sense of Definition 3.3.2), and thus do not violate the no-backtracking principle introduced in Remark 3.2.1.

With these new moment conditions, Assumption 3.3.1 can be shown to be satisfied. Recall that when first introducing Example 1 we assumed π_k is a known measurable function of (Z_k, Y_{-k}) that is linear in parameters θ, and has a gradient (with respect to θ) bounded away from zero for each (z, y_{-k}). We conclude that π_k is Lipschitz in θ, and also satisfies a "reverse Lipschitz" condition; that is, for each (z, y_{-k}) we have:

$$L'_k ||\theta - \theta^*|| \leq |\pi_k(z, y_{-k}; \theta) - \pi_k(z, y_{-k}; \theta^*)| \leq L_k ||\theta - \theta^*||,$$

for some $L'_k, L_k > 0$. Now define:

$$\tau := \min_k \min_{(z,y_{-k})} |0.5 - P(Y_k = 1|Z = z, Y_{-k} = y_{-k})| \quad s.t. \quad |0.5 - P(Y_k = 1|Z = z, Y_{-k} = y_{-k})| > 0. \quad (3.57)$$

Then the analysis in Appendix 3.C.1 shows that Assumption 3.3.1 is verified for $C_1 = L_0 L'$, $C_2 = L_0 L$ and $\delta = \tau/(L_0 L')$, where $L = \min_k L_k$ and $L' = \min_k L'_k$. In Theorem 3.3.1 we can thus take the penalty μ^* to be any value satisfying:

$$\mu^* \geq \max\left\{\frac{L}{L'}, \frac{1}{\tau}\right\}.$$

Theorem 3.3.1 says that the lower and upper envelopes on $I[\varphi](\gamma) = P(Y^\star_\gamma = 1)$, as a function of γ, are given by (3.52) and (3.53), respectively.

Remark 3.3.1 (Counterfactual Coherency)**.** *Recall that Theorem 3.3.1 applies only if the random sets $\boldsymbol{G}^-(\cdot, \theta)$ and $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. In the simultaneous discrete choice example, the counterfactual map $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ can fail to be almost-surely non-empty, which is related to the well known problem of coherency in these models. In particular, for a given instantiation of a vector of unobservables $(u_1, \ldots, u_K)$, there may not exist any vector of counterfactual endogenous outcome variables $(y^\star_{1,\gamma}, \ldots, y^\star_{K,\gamma})$ that solves the system of equations represented by (3.11). However, we note that this issue is unrelated to our particular approach, and might be resolved by (i) conditioning the analysis on the subset of $\mathcal{U}$ that ensures a solution to the system of equations in (3.11), or (ii) imposing certain constraints on the parameter space that ensures the existence of a solution to the system of equations in (3.11). We refer the reader to Chesher and Rosen (2020) for a thorough discussion of this issue. However, whether this "counterfactual coherency" problem can be resolved without violating the no-backtracking principle from Remark 3.2.1 appears to be an open question.*

Example 2 (Program Evaluation (cont'd))**.** *Consider again Example 2 on program evaluation. Verification of Assumption 3.3.1 is presented in Appendix 3.C.2, and uses Lemma 3.3.1 to verify Assumption 3.3.1(ii). Remarkably, we show that Assumption 3.3.1 is satisfied for any value of $\delta > 0$ with $C_1 = C_2 = 1$. Thus we can take the penalty $\mu^* = 1$. Then Theorem 3.3.1 says that the lower and upper envelopes on $I[\varphi](\gamma) = \mathbb{E}[Y^\star_\gamma]$, as a function of γ, are given by (3.52) and (3.53), respectively.*

3.4 On the Learnability of Optimal Policies

In this section, we provide sufficient conditions for PAMPAC learnability. To begin, the following proposition clarifies the connection between the lower envelope function from the previous section and the notion of PAMPAC learnability.

Proposition 3.4.1. *Suppose Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.3.1 hold. Also, suppose that $\varphi : \mathcal{V} \to [\varphi_{lb}, \varphi_{ub}] \subset \mathbb{R}$ is a bounded, measurable function, and that for each $\gamma \in \Gamma$, the random sets $\boldsymbol{G}^-(\cdot, \theta)$ and $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. Then a policy space Γ is PAMPAC learnable with respect to the policy transform of φ if and only if:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P^{\otimes n}_{Y,Z}\left(\sup_{\gamma \in \Gamma} I_{lb}[\varphi](\gamma) - I_{lb}[\varphi](d(\psi)) \leq c\right) \geq \kappa, \quad (3.58)$$

where $I_{lb}[\varphi] : \Gamma \to \mathbb{R}$ is the lower envelope function from Theorem 3.3.1.

Remark 3.4.1. *By Proposition 3.B.1 in Appendix 3.B.2, the map $\psi \mapsto I_{\ell b}[\varphi](d(\psi))$ is universally measurable; that is, measurable with respect to the completion of any $P_{Y,Z} \in \mathcal{P}_{Y,Z}$. Thus, the event in (3.58) can always be assigned a unique probability using outer measures, if necessary.*

In particular, the lower envelope function completely characterizes PAMPAC learnability of the policy space Γ with respect to φ. Thus, it should be unsurprising that our sufficient conditions for a policy space to be PAMPAC learnable will be related to the behaviour of the lower envelope function from Theorem 3.3.1.

Next we introduce an entropy growth condition, which will be imposed as a constraint on the complexity allowed for both the moment functions and the function φ. To introduce the entropy growth condition, we must first define the covering number and metric entropy for a class of functions.

Definition 3.4.1 (Covering Number, Metric Entropy)**.** *Let $(\mathcal{T}, \rho)$ be a semi-metric space. A cover of $\mathcal{T}$ is any collection of sets whose union contains $\mathcal{T}$ as a subset. For any $\varepsilon > 0$, the covering number for $\mathcal{T}$, denoted by $N(\varepsilon, \mathcal{T}, \rho)$, is the smallest number of $\rho-$balls needed to form a $\varepsilon-$cover. The metric entropy is the logarithm of the covering number.*

Definition 3.4.2 (Entropy Growth Condition)**.** *Let $\mathcal{F}$ be a measurable class of real-valued functions on a measurable space $(\mathcal{X}, \mathfrak{A}_{\mathcal{X}})$ with envelope F. The class $\mathcal{F}$ satisfies the entropy growth condition if:*

$$\sup_{Q \in \mathcal{Q}_n} \log N(\varepsilon, \mathcal{F}, ||\cdot||_{Q,2}) = o(n), \tag{3.59}$$

for every $\varepsilon > 0$, with the supremum taken over all discrete probability measures $\mathcal{Q}_n$ on $\mathcal{X}$ with atoms that have probabilities that are integer multiples of $1/n$.

This condition is adapted from a condition in Dudley et al. (1991) (Theorem 6, p. 500) that, in combination with other mild conditions, is shown to be sufficient for a class of functions to be uniform Glivenko-Cantelli.[34] The entropy growth condition essentially says that, for any set $\mathcal{X}_n$ of n points $(x_1, \ldots, x_n)$ in some space $\mathcal{X}$, the logarithm of the minimal number of balls of radius $\varepsilon > 0$ needed to cover the set:

$$\mathcal{F}|_{\mathcal{X}_n} := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n,$$

is of order $o(n)$. Sufficient conditions for this to be the case can be connected to conditions previously used in the literature. For example, (3.59) is satisfied if the class of functions is of VC-type (c.f. Chernozhukov et al. (2013), Belloni et al. (2019)), if the class satisfies Pollard's manageability criterion (c.f. Pollard (1990), Andrews and Shi (2013), Andrews and Shi (2017)), or if the class of functions is otherwise known to be a uniform Donsker class.

The following Theorem shows that if certain classes of functions in the policy analysis problem obey the entropy growth condition, then every policy space is PAMPAC learnable. To state the result, we must first introduce an important class of functions. Let $\Lambda = \{0,1\}^J$, and for a fixed triple $(\theta, \gamma, \lambda) \in \Theta \times \Gamma \times \Lambda$, let $h_{\ell b}(\cdot, \cdot, \theta, \gamma, \lambda) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ be given by:

$$h_{\ell b}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \left(\inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right). \tag{3.60}$$

Note that $h_{\ell b}(\cdot, \cdot, \theta, \gamma, \lambda)$ is exactly the integrand in the lower envelope function from Theorem 3.3.1. Now define the class of functions:

$$\mathcal{H}_{\ell b} := \{h_{\ell b}(\cdot, \cdot, \theta, \gamma, \lambda) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} : (\theta, \gamma, \lambda) \in \Theta \times \Gamma \times \Lambda\}. \tag{3.61}$$

[34] See also Van Der Vaart and Wellner (1996) Theorem 2.8.1 on p.167.

Then we have the following result:

Theorem 3.4.1. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3 and 3.3.1 hold. Also, suppose that* $\varphi : \mathcal{V} \to [\varphi_{\ell b}, \varphi_{ub}] \subset \mathbb{R}$ *is a bounded, measurable function, and that for each* $\gamma \in \Gamma$*, the random sets* $\boldsymbol{G}^-(\cdot, \theta)$ *and* $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ *are almost-surely non-empty for each* $\theta \in \Theta^*$*. Fix any* $\varepsilon > 0$*. (i) If the class of functions* $\mathcal{H}_{\ell b}$ *satisfies the entropy growth condition, then every policy space is PAMPAC learnable with respect to the policy transform of* φ*. Furthermore, for any* $c > 0$ *we have:*

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) \geq c \right) = O(r_1(n)), \tag{3.62}$$

where:

$$r_1(n) := \max \left\{ n^{-1/2}, n^{-1/2} \sup_{Q \in \mathcal{Q}_n} \sqrt{\log N(\varepsilon, \mathcal{H}_{\ell b}, ||\cdot||_{Q,2})} \right\}. \tag{3.63}$$

(ii) If the class of functions:

$$\Phi := \{\varphi(\cdot, u, y^\star) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} : (u, y^\star) \in \mathcal{U} \times \mathcal{Y}^\star\}, \tag{3.64}$$

$$\mathcal{M}_j := \{m_j(\cdot, u, \theta) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} : (u, \theta) \in \mathcal{U} \times \Theta\}, \quad j = 1, \ldots, J, \tag{3.65}$$

are uniformly bounded, and satisfy the entropy growth condition, then so does $\mathcal{H}_{\ell b}$*. Furthermore, for any* $c > 0$ *we have:*

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) \geq c \right) = O(r_2(n)), \tag{3.66}$$

where:

$$r_2(n) := \max \left\{ n^{-1/2}, n^{-1/2} \sup_{Q \in \mathcal{Q}_n} \sqrt{\log N(\varepsilon/4, \Phi, ||\cdot||_{Q,2}) + \sum_{j=1}^{J} \log N(\varepsilon/2, \mathcal{M}_j, ||\cdot||_{Q,2})} \right\}. \tag{3.67}$$

Proof. See Appendix 3.B. ■

The proof of the part (i) proceeds by proposing a specific decision procedure, and then showing that the proposed decision procedure satisfies the requirements of PAMPAC learnability from Definition 3.2.4 when the class of functions $\mathcal{H}_{\ell b}$ satisfies the entropy growth condition. The specific decision procedure proposed in the proof is any procedure that obtains within ε of the maximum of the sample analog lower envelope function for each sample $\psi \in \Psi_n$, for some $\varepsilon > 0$. We call this rule the *$\varepsilon-$maximin empirical rule*, and we will revisit it's properties in the next subsection. Here we also finally see the close connection between PAMPAC learnability and the lower envelope function from Theorem 3.3.1 in the previous section, which has been alluded to throughout the paper. The particular form of the lower envelope function from Theorem 3.3.1 makes it amenable to analysis using methods from empirical process theory, which are used in the proof of Theorem 3.4.1. Also note that Assumption 3.3.1, which was needed to obtain a bound on the penalty μ^* in Theorem 3.3.1, is also needed for this result. Without a bound on this penalty, Theorem 3.4.1 will generally not be true.

The proof of part (ii) of Theorem 3.4.1 shows that if each "component" of the lower envelope of the policy transform—namely the moment functions and the function φ—satisfy the entropy growth condition, then the metric entropy of the class $\mathcal{H}_{\ell b}$ can also be controlled. Combined with the result in Proposition 3.4.1, the proof of part (ii) of Theorem 3.4.1 then shows that our proposed $\varepsilon-$maximin decision rule can obtain

close to the maximum value (over $\gamma \in \Gamma$) of the lower envelope of the policy transform with high probability.

It may seem surprising that our learnability result holds for any policy space. However, this is a result of the fact that the complexity of the policy space is tempered by the class of functions Φ from (3.64), since it is only through functions in this class that the policy can affect the policy transform. By imposing that the class Φ satisfy the entropy growth condition, we are implicitly imposing constraints on the complexity of the policy space. Note that the Theorem provides only sufficient conditions for PAMPAC learnability, and alternative results that impose complexity constraints on the policy space Γ directly, rather than on Φ, may be possible.

We will now turn to our motivating examples to verify learnability of the involved policy spaces.

Example 1 (Simultaneous Discrete Choice (cont'd))**.** *Consider again Example 1 on simultaneous discrete choice. In this case we have:*

$$\Phi := \{\mathbb{1}\{\pi_k(\gamma(\,\cdot\,);\theta) \geq u\} : (u,\theta) \in \mathcal{U} \times \Theta\}, \tag{3.68}$$

with the moment conditions:

$$\mathbb{E}\left[\left(\mathbb{1}\{U_k \leq \pi_k(z', y'_{-k};\theta)\} - \max\{L_0\pi_k(z', y'_{-k};\theta), 0\} - 0.5\right)\mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}\right] \leq 0, \tag{3.69}$$

$$\mathbb{E}\left[\left(0.5 - \mathbb{1}\{U_k \leq \pi_k(z', y'_{-k};\theta)\} - \max\{-L_0\pi_k(z', y'_{-k};\theta), 0\}\right)\mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\}\right] \leq 0, \tag{3.70}$$

for $k = 1, \ldots, K$, for all $z, z' \in \mathcal{Z}$ and all $y_{-k}, y'_{-k} \in \mathcal{Y}^{K-1}$. Details on the verification of the entropy growth condition for both Φ and the class of moment functions associated with the moment conditions above are presented in Appendix 3.C.1. Furthermore, under our assumptions for this example, the rate of convergence derived from Theorem 3.4.1 is found to be $O(n^{-1/2})$.

Example 2 (Program Evaluation (cont'd))**.** *Consider again Example 2 on program evaluation. In this case we have:*

$$\Phi := \{\mathbb{1}\{g(\gamma(z)) \geq u\}(u_1 - u_0) + u_0 : (u_0, u_1, u, g) \in \mathcal{U} \times \mathcal{G}\}, \tag{3.71}$$

with the moment conditions:

$$\mathbb{E}[(D - g(Z_0, X))\,\mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.72}$$

$$\mathbb{E}[(g(Z_0, X) - D)\,\mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.73}$$

$$\mathbb{E}[(\mathbb{1}\{U \leq g(z_0, x)\} - g(z_0, x))\,\mathbb{1}\{X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.74}$$

$$\mathbb{E}[(g(z_0, x) - \mathbb{1}\{U \leq g(z_0, x)\})\,\mathbb{1}\{X = x\}] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X}, \tag{3.75}$$

$$\mathbb{E}\left[t(z_0, x) - \mathbb{1}\{Z = z_0, X = x\}\right] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, \forall x \in \mathcal{X}, \tag{3.76}$$

$$\mathbb{E}\left[\mathbb{1}\{Z = z_0, X = x\} - t(z_0, x)\right] \leq 0, \qquad \forall z_0 \in \mathcal{Z}_0,\, \forall x \in \mathcal{X}, \tag{3.77}$$

and:

$$\mathbb{E}\left[U_d\left(\mathbb{1}\{Z = z_0, X = x\}\sum_{z_0 \in \mathcal{Z}_0} t(z_0, x) - \mathbb{1}\{X = x\}t(z_0, x)\right)\right] \leq 0,\ \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X},\, d \in \{0, 1\}, \tag{3.78}$$

$$\mathbb{E}\left[U_d\left(\mathbb{1}\{X = x\}t(z_0, x) - \mathbb{1}\{Z = z_0, X = x\}\sum_{z_0 \in \mathcal{Z}_0} t(z_0, x)\right)\right] \leq 0,\ \forall z_0 \in \mathcal{Z}_0,\, x \in \mathcal{X},\, d \in \{0, 1\}. \tag{3.79}$$

Details on the verification of the entropy growth condition for both Φ and the class of functions associated

with the moment functions above are presented in Appendix 3.C.2. Furthermore, under our assumptions for this example, the rate of convergence derived from Theorem 3.4.1 is found to be $O(n^{-1/2})$.

3.5 Ex-Post Theoretical Results

Theorem 3.4.1 shows sufficient conditions for PAMPAC learnability in a given environment. However, while the result shows that it may be possible *ex-ante* (i.e. before observing a particular sample) to learn a given policy space, it does not provide us any useful *ex-post* (i.e. after observing the sample) information on the performance of our decision rule. This reflects a well-known complaint of PAC learnability, and has given rise to the literature on data-dependent excess risk bounds in statistical learning literature; see Bartlett et al. (2002), Koltchinskii (2001), and Koltchinskii (2006) for examples, and Boucheron et al. (2005) or Koltchinskii (2011) for a review. Thus after establishing learnability of a particular class of policies, it may be of separate interest to evaluate the finite sample performance of a given decision rule for a given sample.

This is accomplished in the next subsections. We will focus our attention on the particular decision rule used in the proof of Theorem 3.4.1 which was shown to satisfy the requirements of PAMPAC learnability under the assumptions of the theorem. The decision rule used was allowed to be any $\varepsilon-$maximizer of the empirical version of the lower envelope function $I_{\ell b}[\varphi](\gamma)$, which is why we will call it the $\varepsilon-$*maximin empirical rule.*

Definition 3.5.1 ($\varepsilon-$maximin empirical welfare)**.** *Fix any* $\varepsilon \geq 0$ *and let* $\widehat{I}_{\ell b}[\varphi](\gamma)$ *denote the lower envelope from Theorem 3.3.1 evaluated at the empirical measure for* (Y, Z)*. Then* $d : \Psi_n \to \Gamma$ *is a* $\varepsilon-$*maximin empirical (eME) rule if:*

$$\widehat{I}_{\ell b}[\varphi](d(\psi)) + \varepsilon \geq \sup_{\gamma \in \Gamma} \widehat{I}_{\ell b}[\varphi](\gamma). \tag{3.80}$$

Remark 3.5.1. *Note that in general the "*ε*" is necessary (although it can be made arbitrarily small), owing to the fact that the supremum of* $\widehat{I}_{\ell b}[\varphi](\cdot)$ *may not be obtained.*

Furthermore, unlike our result on PAMPAC learnability, all of the results in the next subsections are data-dependent, and do not depend on any particular properties (beyond measurability) of any function classes involved in the policy decision problem. Thus, there is no need to verify the entropy growth condition, or any other condition sufficient for learnability to use the results ahead. In practice, we still recommend that the sufficient conditions for learnability of a policy space be verified prior to using the results.

3.5.1 Theoretical Results for the Maximin Empirical Rule

In this section we obtain a bound on the value of $c_n(d, \kappa)$ for any fixed κ taking d to be the eME rule. To describe our procedure, we will first introduce a data-dependent complexity measure for the class $\mathcal{H}_{\ell b}$. The complexity measure we use is based on the empirical Rademacher complexity, advocated by Bartlett et al. (2002), Koltchinskii (2001), and Koltchinskii (2006) (among others) in the context of empirical risk minimization.

Definition 3.5.2 (Empirical Rademacher Complexity)**.** *Let* $\mathcal{F}$ *be a class of measurable functions* $f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$*. The empirical Rademacher complexity of* $\mathcal{F}$ *is given as:*

$$||\mathfrak{R}_n||(\mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \cdot f(y_i, z_i) \right|, \tag{3.81}$$

where ξ_i are realizations of Rademacher random variables; that is, $\xi \in \{-1, 1\}$ and $P(\xi_i = -1) = P(\xi = 1) = 1/2$.

Remark 3.5.2. *A technical point worth emphasizing is that, when seen as a function of the underlying product probability space, the empirical Rademacher complexity may not be a measurable function. We suppress these difficulties in the statement of our results, although we show in Appendix 3.B.2 that the Rademacher complexity $||\mathfrak{R}_n||(\mathcal{H}_{\ell b})$ is universally measurable (with respect to the product Borel $\sigma-$algebra on $(\mathcal{Y} \times \mathcal{Z})^n$), which is sufficient for the purposes in this paper.*

In our context, the empirical Rademacher complexity of the class $\mathcal{H}_{\ell b}$ depends only on the observed empirical distribution and on n draws of a Rademacher random variable; it can therefore be computed after simulating from the Rademacher distribution. With this new definition in hand, we have the following result:

Theorem 3.5.1. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.3.1 hold. Let $\varphi : \mathcal{V} \to [\varphi_{\ell b}, \varphi_{ub}] \subset \mathbb{R}$ be a bounded, measurable function, and suppose that for each $\gamma \in \Gamma$, the random sets $\boldsymbol{G}^-(\cdot, \theta)$ and $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. Let $\{(y_i, z_i)\}_{i=1}^n$ be i.i.d. from some distribution $P_{Y,Z}$ satisfying our assumptions and let $d : \Psi_n \to \Gamma$ be an eME decision rule for some $\varepsilon > 0$. Furthermore, let $\overline{H} < \infty$ satisfy $|h| \leq \overline{H}$ for every $h \in \mathcal{H}_{\ell b}$, and let:*

$$c_n(\kappa) = 4||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) + \sqrt{\frac{72 \ln(2/(2-\kappa))\overline{H}^2}{n}} + 5\varepsilon. \tag{3.82}$$

Then for any sample size n, and any $\kappa \in (0, 1)$ we have:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) \leq c_n(\kappa) \right) \geq \kappa. \tag{3.83}$$

Proof. See Appendix 2.A. ■

Theorem 3.5.1 shows two closely related results. First, for any fixed value of $\kappa \in (0, 1)$ the Theorem shows that, when in the worst-case state, the eME rule obtains within $c_n(\kappa)$ of the maximin value of the state-dependent policy transform with probability at least κ. Simple comparative statics show that the value of $c_n(\kappa)$ is smaller when n is larger and/or $||\mathfrak{R}_n||(\mathcal{H}_{\ell b})$ and $\overline{H}$ are smaller. The only difficult part of computing $c_n(\kappa)$ is computing the Rademacher complexity, which is approximately as difficult computationally as computing the empirical version of the lower bound in Theorem 3.3.1.

We again see a close connection between PAMPAC learnability and the lower envelope function from Theorem 3.3.1. The particular form of the lower envelope function from Theorem 3.3.1 makes it especially amenable to analysis using concentration concentration inequalities, which are used in the proof of Theorem 3.5.1. Again Assumption 3.3.1 is required for this result: without a finite (and known) value for the penalty μ^*, derivation of the finite sample results in Theorem 3.5.1 would not be possible.

Finally we mention again that, unlike Theorem 3.4.1 on PAMPAC learnability, Theorem 3.5.1 does not impose any restrictions on the underlying class of functions $\mathcal{H}_{\ell b}$. In particular, this class need not satisfy the entropy growth condition from Definition 3.4.2, nor any other sufficient conditions for learnability, meaning Theorem 3.5.1 is applicable even when Γ is not PAMPAC learnable. As a result, Theorem 3.5.1 is able to provide finite sample guarantees for the eME rule, but necessarily remains silent about rates of convergence.

3.5.2 Bounds on the Set of Optimal Policies

The previous subsection uses a specific rule, the eME rule, and derives finite sample theoretical guarantees on the performance of this rule. However, the eME rule is only one particular rule, and for a variety of

reasons it may not be the rule selected by the policymaker.

In order to complement the results of the previous subsection, in this subsection we will provide some theoretical results on alternative policy rules. To understand the approach, let us define the function:

$$\mathscr{E}^*(\gamma) := \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) = \sup_{\gamma \in \Gamma} I_{\ell b}[\varphi](\gamma) - I_{\ell b}[\varphi](\gamma), \tag{3.84}$$

and the set:

$$\mathscr{G}^*(\delta) := \{\gamma \in \Gamma : \mathscr{E}^*(\gamma) \leq \delta\}. \tag{3.85}$$

We call the set $\mathscr{G}^*(\delta)$ the $\delta-$level set. Our objective in this subsection will be to provide an approximation of the $\delta-$level set that holds with probability at least κ. If we can do so, then by constructionn any decision rule $d : \Psi_n \to \Gamma$ that maps within our approximation of the $\delta-$level set will have $c_n(d, \kappa) \leq \delta$. There may be many decision rules that map within our approximation to the $\delta-$level set, so our theoretical results will be applicable to a large number of decision rules. As a by product of our analysis, we will also show that for certain values of δ the eME rule will be contained in the $\delta-$level set with probability at least κ. Again, the results of this section do not impose any restrictions on the underlying class of functions $\mathcal{H}_{\ell b}$, and are applicable even when Γ is not PAMPAC learnable.

To introduce our results for the $\delta-$level set, we must first introduce some additional notation. In particular, define:

$$\mathscr{E}_n(\gamma) := \sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} \widehat{I}[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} \widehat{I}[\varphi](\gamma, s) = \sup_{\gamma \in \Gamma} \widehat{I}_{\ell b}[\varphi](\gamma) - \widehat{I}_{\ell b}[\varphi](\gamma), \tag{3.86}$$

and for $\delta > 0$ define the set:

$$\mathscr{G}_n(\delta) := \{\gamma \in \Gamma : \mathscr{E}_n(\gamma) \leq \delta\}. \tag{3.87}$$

The set $\mathscr{G}_n(\delta)$ represents the empirical version of the $\delta-$level set.

The following theorem shows that, for sufficiently large δ, the $\delta-$level set is contained within an enlargement of, and contains a contraction of, the empirical $\delta-$level set with high probability.

Theorem 3.5.2. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.3.1 hold. Also suppose that $\varphi : \mathcal{V} \to [\varphi_{\ell b}, \varphi_{ub}] \subset \mathbb{R}$ is a bounded, measurable function, and that for each $\gamma \in \Gamma$, the random sets $\boldsymbol{G}^-(\cdot, \theta)$ and $\boldsymbol{G}^\star(\cdot, \theta, \gamma)$ are almost-surely non-empty for each $\theta \in \Theta^*$. Let $\overline{H} < \infty$ satisfy $|h| \leq \overline{H}$ for every $h \in \mathcal{H}_{\ell b}$, and suppose that $\{(y_i, z_i)\}_{i=1}^n$ is i.i.d. from some distribution $P_{Y,Z}$ satisfying our assumptions. Define:*

$$\mathcal{H}'_{n,\ell b}(\delta) := \{h_{\ell b}(\cdot, \cdot, \theta, \gamma, \lambda) - h_{\ell b}(\cdot, \cdot, \theta', \gamma', \lambda') : \theta, \theta' \in \Theta,\ \gamma, \gamma' \in \mathscr{G}_n(\delta), \lambda, \lambda' \in \{0, 1\}^J\},$$

where $\mathcal{H}'_{n,\ell b}(\delta)$ has a uniform bound $\overline{H}'_n(\delta) \leq 2\overline{H} < \infty$. Furthermore, let $t_j := \sqrt{c_1 \log(c_2 j)}$ with $c_1 = 5$ and $c_2 = (3/(2(1-\kappa)))^{2/5}$, and let $\{\delta_j\}_{j=0}^\infty$ be a sequence decreasing to zero with $\delta_0 > 2\overline{H}$. Choose some $\mathfrak{a} \in (1, \infty)$, let $\mathfrak{b} = 2 - 1/\mathfrak{a}$, and let:

$$T_n(\delta) := \begin{cases} 2||\mathfrak{R}_n||(\mathcal{H}'_{n,\ell b}(\mathfrak{b}\delta_j)) + \frac{3t_j \overline{H}'_n(\mathfrak{b}\delta_j)}{\sqrt{n}}, & \text{if } \delta \in (\delta_{j+1}, \delta_j] \text{ for some } j \geq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{3.88}$$

and:

$$T_n^{\flat}(\sigma) := \sup_{\delta \geq \sigma} \frac{T_n(\delta)}{\delta}, \tag{3.89}$$

$$T_n^{\sharp}(\eta) := \inf \left\{ \sigma > 0 : T_n^{\flat}(\sigma) \leq \eta \right\}. \tag{3.90}$$

Finally, set $\delta^* > T_n^{\sharp}(1 - 1/\mathfrak{a})$. *Then for any* $\delta \geq \mathfrak{a}\delta^*$ *we have:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\mathscr{G}_n(\delta/\mathfrak{a}) \subseteq \mathscr{G}^*(\delta) \subseteq \mathscr{G}_n(\mathfrak{b}\delta) \right) \geq \kappa.$$

Proof. See Appendix 3.B. ■

Theorem 3.5.2 closely mimics results in the statistical learning literature, namely in the problem of bounding excess risk in empirical risk minimization problems. In particular, the proof of the result uses techniques developed by Koltchinskii (2006) and Koltchinskii (2011), where the latter gives a textbook treatment.[35] Theorem 3.5.2 gives a novel application of these techniques to the problem of policy choice in the presence of partial identification. Similar to the other results in this paper, Theorem 3.5.2 relies crucially on the form of the lower envelope function from Theorem 3.3.1. Again Assumption 3.3.1 is required, since Theorem 3.5.2 requires a finite (and known) value for the penalty parameter μ^*.

Intuitively, Theorem 3.5.2 says that for a suitably large value of δ the $\delta-$level sets $\mathscr{G}_n(\delta)$ of the function $\mathscr{E}_n(\cdot)$ can be used to approximate the $\delta-$level sets $\mathscr{G}^*(\cdot)$ of the function $\mathscr{E}^*(\cdot)$. The substantial component of the results is the selection of such a "suitably large value of δ." In particular, the value of δ needed for our approximation to work must be larger than the value of δ^* from the Theorem, where δ^* is related to the solution of a fixed point equation. The connection of the functions $T_n(\,\cdot\,)$, $T_n^{\flat}(\,\cdot\,)$ and $T_n^{\sharp}(\,\cdot\,)$ to fixed point equations is illustrated in Figure 3.4 and is described in its associated caption. As illustrated in the Figure, the function $T_n(\delta)$ is a left-continuous step function that is greater than or equal to zero on the interval $[0, \delta_0]$, and zero otherwise.

The proof of Theorem 3.5.2 relies on Lemma 3.5.1, and the best way to understand Theorem 3.5.2 is to first understand Lemma 3.5.1.

Lemma 3.5.1. *Suppose that the assumptions of Theorem 3.5.2 all hold. Define:*

$$\mathcal{H}'_{\ell b}(\delta) := \{h_{\ell b}(\cdot,\cdot,\theta,\gamma,\lambda) - h_{\ell b}(\cdot,\cdot,\theta',\gamma',\lambda') : \theta,\theta' \in \Theta,\ \gamma,\gamma' \in \mathscr{G}^*(\delta), \lambda,\lambda' \in \Lambda\},$$

where $\mathcal{H}'_{\ell b}(\delta)$ *has a uniform bound* $\overline{H}'(\delta) \leq 2\overline{H} < \infty$. *Furthermore, let* $t_j := \sqrt{c_1 \log(c_2 j)}$ *with* $c_1 = 5$ *and* $c_2 = (3/(2(1-\kappa)))^{2/5}$, *and let* $\{\delta_j\}_{j=0}^{\infty}$ *be a sequence decreasing to zero with* $\delta_0 > 2\overline{H}$. *Also, let:*

$$T(\delta) := \begin{cases} 2||\mathfrak{R}_n||(\mathcal{H}'_{\ell b}(\delta_j)) + \frac{3t_j \overline{H}'(\delta_j)}{\sqrt{n}}, & \text{if } \delta \in (\delta_{j+1}, \delta_j], \\ 0, & \text{otherwise,} \end{cases} \tag{3.91}$$

and:

$$T^{\flat}(\sigma) := \sup_{\delta \geq \sigma} \frac{T(\delta)}{\delta}, \tag{3.92}$$

$$T^{\sharp}(\eta) := \inf \left\{ \sigma > 0 : T^{\flat}(\sigma) \leq \eta \right\}. \tag{3.93}$$

[35] The $\flat-$ and $\sharp-$transforms are taken from Koltchinskii (2006), and the properties of these transforms can be found in Appendix A.3. of Koltchinskii (2011).

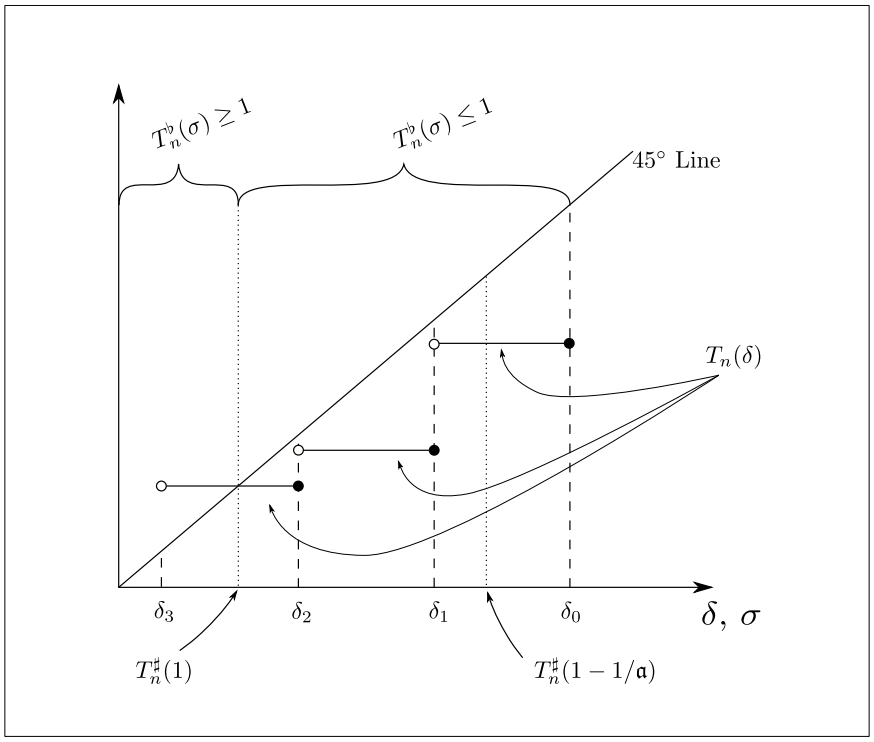


Figure 3.4: This figure illustrates step (iv) in the procedure to determine the $\delta-$level set. After choosing a decreasing sequence $\{\delta_j\}_{j=0}^{\infty}$, the policymaker finds the value δ^* such that $\delta^* > T_n^\sharp(1-1/\mathfrak{a})$. In the figure, this occurs in the interval $(\delta_1, \delta_0]$ (although, of course, this need not be the case). The figure also illustrates the fact that $T_n(\delta)$ is a step function. Finally, the figure illustrates how the $\flat-$ and $\sharp-$transforms of $T_n(\delta)$ are related to fixed point-equations. In particular, the figure illustrates the fixed point of $T_n(\delta) = \delta$, which is given exactly by $T_n^\sharp(1)$. In addition, the fixed point of $T_n(\delta) = \delta(1-1/\mathfrak{a})$ is given by $T_n^\sharp(1-1/\mathfrak{a})$.

Finally, suppose $\delta^{**} > T^\sharp(1-1/\mathfrak{a})$ *for some* $\mathfrak{a} \in (1,\infty)$. *Then for any* $\delta \geq \mathfrak{a}\delta^{**}$ *we have:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(\mathscr{G}_n(\delta/\mathfrak{a}) \subseteq \mathscr{G}^*(\delta) \subseteq \mathscr{G}_n((2-1/\mathfrak{a})\delta)\right) \geq \kappa.$$

Proof. See Appendix 3.B. ■

Note that Lemma 3.5.1 is very similar to Theorem 3.5.2, with a major exception being that the class of functions $\mathcal{H}'_{\ell b}(\delta)$ in Lemma 3.5.1 differs from the class of functions $\mathcal{H}'_{n,\ell b}(\delta)$ in Theorem 3.5.2. Note that $\mathcal{H}'_{n,\ell b}(\delta)$ represents a "feasible version" of $\mathcal{H}'_{\ell b}(\delta)$, since $\mathcal{H}'_{\ell b}(\delta)$ depends on the unknown $\delta-$level set $\mathscr{G}^*(\delta)$, where $\mathcal{H}'_{n,\ell b}(\delta)$ depends on the empirical $\delta-$level set $\mathscr{G}_n(\delta)$.

A heuristic proof may help provide some sense of how these results work. A necessary step in proving either Theorem 3.5.2 or Lemma 3.5.1 is to relate the quantities $\mathscr{E}_n(\gamma)$ and $\mathscr{E}^*(\gamma)$, which is exactly what is done in the proof of Lemma 3.5.1. Among other things, the proof of Lemma 3.5.1 demonstrates that an important object connecting the quantities $\mathscr{E}_n(\gamma)$ and $\mathscr{E}^*(\gamma)$ is given by:

$$\delta \mapsto \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\delta)} \sup_{\lambda,\lambda' \in \Lambda} \left| (\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \mathbb{P}_n h_{\ell b}(\cdot,\theta',\gamma',\lambda')) - (P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - P h_{\ell b}(\cdot,\theta',\gamma',\lambda')) \right|, \tag{3.94}$$

where:

$$\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) := \frac{1}{n}\sum_{i=1}^{n} h_{\ell b}(y_i,z_i,\theta,\gamma,\lambda), \qquad P h_{\ell b}(\cdot,\theta,\gamma,\lambda) := \int h_{\ell b}(y,z,\theta,\gamma,\lambda)\, dP_{Y,Z}.$$

The quantity (3.94) is easily seen to be the sup-norm of a particular empirical process. Note that this empirical process depends on unknown population quantities through both $\mathscr{G}^*(\delta)$ and through the functions $Ph_{\ell b}(\cdot,\theta,\gamma,\lambda)$ and $Ph_{\ell b}(\cdot,\theta',\gamma',\lambda')$, which depend on the unknown true probability measure. While the dependence on $\mathscr{G}^*(\delta)$ is unavoidable for now, the dependence on $Ph_{\ell b}(\cdot,\theta,\gamma,\lambda)$ and $Ph_{\ell b}(\cdot,\theta',\gamma',\lambda')$ can be removed by working with the function $T(\delta)$ from (3.91).[36] Thus, the function $T(\delta)$ in Lemma 3.5.1—which is slightly different from $T_n(\delta)$ in Theorem 3.5.2—is constructed to serve as an upper envelope of the quantity in (3.94), for every $\delta \in [0,\delta_0]$, on some event E_n with probability at least κ.

With (3.94) replaced by its upper bound $T(\delta)$, the proof of Lemma 3.5.1 then shows that, if $\sigma := \mathscr{E}^*(\gamma)$, the following inequalities hold on the event E_n:

$$\mathscr{E}^*(\gamma) \leq \mathscr{E}_n(\gamma) + T(\sigma), \tag{3.95}$$

$$\mathscr{E}_n(\gamma) \leq \mathscr{E}^*(\gamma) + T(\sigma). \tag{3.96}$$

Now note that if $\delta^{**} = T^\sharp(1-1/\mathfrak{a})+\varepsilon$ for any $\varepsilon > 0$, then $T(\delta) \leq (1-1/\mathfrak{a})\cdot\delta$ for every $\delta \geq \delta^{**}$. Furthermore, by construction the value of δ^{**} will be close to the smallest possible value for which this is true. Now fix any γ with $\sigma = \mathscr{E}^*(\gamma) \geq \delta^{**}$. Then clearly:

$$T(\sigma) \leq \left(1 - \frac{1}{\mathfrak{a}}\right) \mathscr{E}^*(\gamma), \tag{3.97}$$

Combining this result with (3.95) and (3.96) we obtain that for any γ satisfying $\mathscr{E}^*(\gamma) \geq \delta^{**}$ we have:

$$\mathscr{E}^*(\gamma) \leq \mathfrak{a}\mathscr{E}_n(\gamma), \tag{3.98}$$

$$\mathscr{E}_n(\gamma) \leq \mathfrak{b}\mathscr{E}^*(\gamma). \tag{3.99}$$

The remainder of the proof of Lemma 3.5.1 is dedicated to showing that the following inequalities hold for any $\gamma \in \Gamma$ on the event E_n:[37]

$$\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right), \tag{3.100}$$

$$\mathscr{E}_n(\gamma) \leq \mathfrak{b}\left(\mathscr{E}^*(\gamma) \vee \delta^{**}\right), \tag{3.101}$$

After these inequalities are established, it is straightforward to argue that $\mathscr{G}_n(\delta/\mathfrak{a}) \subseteq \mathscr{G}^*(\delta) \subseteq \mathscr{G}_n(\mathfrak{b}\delta)$ on the event E_n when $\delta \geq \mathfrak{a}\delta^{**}$. Intuitively, the proof of Theorem 3.5.2 then shows that $\mathcal{H}'_{\ell b}(\delta)$, $T(\cdot)$ (and its $\flat-$ and $\sharp-$transform) and δ^{**} defined in Lemma 3.5.1 can be replaced with their feasible versions $\mathcal{H}'_{n,\ell b}(\delta)$, $T_n(\cdot)$ (and its $\flat-$ and $\sharp-$transform) and δ^* defined in Theorem 3.5.2.

Theorem 3.5.2 suggests the following procedure to approximate the $\delta-$level set. The policymaker begins by computing $\mathscr{E}_n(\gamma)$ as a function of γ (for example, by establishing a grid over Γ). The policymaker fixes some value $\mathfrak{a} \in (1,\infty)$ and constructs a sequence $\{\delta_j\}_{j=0}^{\infty}$ decreasing to zero with $(1-1/\mathfrak{a})\delta_0 > 2\overline{H}$. In general the procedure will give a tighter bound if the sequence $\{\delta_j\}_{j=0}^{\infty}$ has small initial increments. The policymaker then computes $\delta^* > T_n^\sharp(1-1/\mathfrak{a})$. This is done by the following procedure:

(i) The policymaker takes n i.i.d. draws of a Rademacher random variable ξ.

(ii) At the j^{th} step (beginning at step 0) the policymaker uses $\mathscr{E}_n(\gamma)$ to compute the Rademacher complexity

[36] Note that, technically speaking, the dependence of (3.94) on $Ph_{\ell b}(\cdot,\theta,\gamma,\lambda)$ and $Ph_{\ell b}(\cdot,\theta',\gamma',\lambda')$ is removed using a symmetrization inequality (c.f. Van Der Vaart and Wellner (1996) Lemma 2.3.1) and a Hoeffding-type concentration inequality, which leads exactly to the upper-bound $T(\delta)$, which holds with high probability.

[37] Note that the first inequality is trivial, since (3.98) is satisfied when $\mathscr{E}^*(\gamma) \geq \delta^{**}$, and if $\mathscr{E}^*(\gamma) \leq \delta^{**}$, then $\mathscr{E}^*(\gamma) \leq \mathfrak{a}\delta^{**}$, since $\mathfrak{a} > 1$. The second of these inequalities is non-trivial, and relies on an auxiliary result given by Lemma 3.B.9 in the Appendix.

$||\mathfrak{R}_n||(\mathcal{H}'_{n,\ell b}(\mathfrak{b}\delta_j))$ with the formula (3.81).

(iii) The policymaker uses $\mathscr{E}_n(\gamma)$ to compute a uniform upper bound $\overline{H}_n(\delta_j)$ for $\mathcal{H}'_{n,\ell b}(\delta_j)$ (or she can simply use $2\overline{H}$).

(iv) The policymaker determines if there is any value $\delta \in (\delta_{j+1}, \delta_j]$ such that $T_n(\delta_j)/\delta \geq 1 - 1/\mathfrak{a}$.

- If so, the policymaker stops and sets $\delta^* = \delta + \eta$, where $\eta > 0$ and $\delta \in (\delta_{j+1}, \delta_j]$ is equal to any value satisfying $T_n(\delta_j)/\delta \leq 1 - 1/\mathfrak{a}$.
- If not, the policymaker repeats steps (i) and (ii) for iteration $j + 1$.

An illustration of this step is provided in Figure 3.4. By Theorem 3.5.2, the policymaker then knows that for every $\delta \geq \delta^*$, the $\delta-$minimal set $\mathscr{G}(\delta)$ will be contained within the sample analogue $\delta-$minimal set $\mathscr{G}_n(\mathfrak{b}\delta)$, and will contain the sample analogue $\delta-$minimal set $\mathscr{G}_n(\delta/\mathfrak{a})$ with probability at least κ. Note that the computational bottleneck in this procedure arises from repeatedly computing the Rademacher complexity.

In addition to being interesting in its own right, Theorem 3.5.2 also sheds light on the results from the previous subsection. In particular, the proof of Theorem 3.5.2 and Lemma 3.B.9 lead to the following result, which is stated as a corollary of Theorem 3.5.2.

Corollary 3.5.1. *Suppose the assumptions of Theorem 3.5.2 hold, and let δ^* be as in Theorem 3.5.2. For any $\varepsilon > 0$ let $\hat{\gamma} \in \Gamma$ be the policy selected by the eME decision rule. If $\delta \geq \delta^* \geq \varepsilon > 0$, then:*

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\mathscr{E}^*(\hat{\gamma}) \leq \delta \right) \geq \kappa.$$

That is $\hat{\gamma} \in \mathscr{G}^*(\delta)$ with high probability when $\delta \geq \delta^* \geq \varepsilon > 0$.

This result shows that, if $\varepsilon \leq \delta^*$ then our eME rule from the previous subsection will be contained in the $\delta-$level set $\mathscr{G}^*(\delta)$ when $\delta \geq \delta^*$ with high probability. This should serve as some additional justification for using the eME rule, since it shows that, when both δ^* and ε are small, the procedure suggested by Theorem 3.5.2 will not lead to decision rules that vastly outperform the eME rule.

3.6 Conclusion

The purpose of the paper is to develop a general and novel framework for bounding counterfactual quantities and for making policy decisions. Our framework is applicable in models that partially identified and/or incomplete. Furthermore, we do not require parametric distributional assumptions for the latent variables, and we allow for moment conditions that depend on latent variables. We introduce the *policy transform*, and argue that many counterfactual quantities can be written as the policy transform of some function. We then introduce a preference relation that respects weak dominance, and discuss the problem of policy choice using a framework similar to the PAC model of learnability from computational learning theory. Our theoretical results are divided into those that are applicable ex-ante (i.e. before observing the sample) and ex-post (i.e. after observing the sample). For our ex-ante results, we introduce the notion of "learning" a policy space, and provide sufficient conditions for a policy space to be learnable. For our ex-post results, we provide theoretical guarantees on the performance of particular policy rules. Throughout the paper we also demonstrate how to apply the results to a simultaneous discrete choice example and a program evaluation example.

There are many obvious extensions of this work that might be interesting. This paper has been particularly focused on theoretical developments, with examples serving mainly a pedagogical purpose. Further

development of the examples and empirical applications are needed to clearly illustrate and fully investigate the strengths and weaknesses of the method in practice. In addition, the paper has been largely silent on implementation, which may be computationally complex in certain environments. Further development of efficient algorithms to implement the procedures is clearly needed. Finally, the relation between PAC learnability and the literature on frequentist decision theory requires further investigation and clarification. We believe all of these extensions to be fruitful avenues of future research.

Appendix 3.A Preliminaries

3.A.1 Preliminaries on Random Set Theory

This Appendix introduces some key elements of random set theory. Since measurability issues play a significant role in random set theory, we begin by providing the definition of an Effros-measurable multifunction, and show its connection with the definition of a random set.

Definition 3.A.1 (Effros-Measurability, Random Set)**.** *Let $(\Omega, \mathfrak{A}, P)$ be a probability space, let $\mathcal{V}$ be a Polish space, and let $\mathcal{O}_{\mathcal{V}}$ denote the collection of all open sets on $\mathcal{V}$. A multifunction $\boldsymbol{V} : \Omega \to \mathfrak{F}_{\mathcal{V}}$ is called Effros-measurable if for every $A \in \mathcal{O}_{\mathcal{V}}$ we have $\boldsymbol{V}^-(A) := \{\omega \in \Omega : \boldsymbol{V}(\omega) \cap A \neq \emptyset\} \in \mathfrak{A}$. An Effros-measurable closed-valued multifunction on a probability space $(\Omega, \mathfrak{A}, P)$ is called a random closed set.*

From this definition, we see that a random closed set is an Effros-measurable closed multifunction which takes elements from the underlying probability space to the collection of closed sets on some Polish space $\mathcal{V}$. An Effros-measurable closed multifunction is also sometimes called *weakly measurable*.[38] When the underlying probability space $(\Omega, \mathfrak{A}, P)$ is complete Effros-measurability is equivalent to both (i) $\boldsymbol{V}^-(B) \in \mathfrak{A}$ for all $B \in \mathfrak{B}(\mathcal{V})$ (Borel measurability) and (ii) $\boldsymbol{V}^-(F) \in \mathfrak{A}$ for all $F \in \mathfrak{F}_{\mathcal{V}}$ (strong measurability).[39] Our main interest in the paper is in the case when $\mathcal{V}$ is a subset of finite-dimensional euclidean space, although the framework is more general.

While Effros-measurability is the proper notion of measurability for many of the results, it can be difficult to verify. There are other conditions that are sufficient for Effros measurability, but we find one condition to be particularly helpful in the examples. Let d denote the metric on a Polish space $\mathcal{V}$, and let $\boldsymbol{V} : \Omega \to \mathfrak{F}_{\mathcal{V}}$ be a multifunction. The distance to the set $\boldsymbol{V}(\omega)$ on $\mathcal{V}$ is given by:

$$d(v, \boldsymbol{V}(\omega)) := \inf\{d(v, v') : v' \in \boldsymbol{V}(\omega)\}.$$

By a result of Himmelberg (1975), Effros measurability of the multifunction $\boldsymbol{V}$ is equivalent to measurability of $d(v, \boldsymbol{V}(\omega))$ (as a random variable from Ω to $[0, \infty]$) for each $v \in \mathcal{V}$.

Throughout the paper it is also important to understand what it means for two random sets to be identically distributed, which is provided in the next definition.

Definition 3.A.2 (Identically Distributed Random Sets)**.** *Let $(\Omega, \mathfrak{A}, P)$ be a probability space, let $\mathcal{V}$ be a Polish space. We say that two random sets $\boldsymbol{V}$ and $\boldsymbol{V}^*$ are identically distributed, denoted by $\boldsymbol{V} \sim \boldsymbol{V}^*$, if for every $A \in \mathcal{O}_{\mathcal{V}}$ we have $P(\omega : \boldsymbol{V}(\omega) \cap A \neq \emptyset) = P(\omega : \boldsymbol{V}^*(\omega) \cap A \neq \emptyset)$.*

Finally, an important concept in random set theory is that of a selection from a random set. Intuitively, a random set $\boldsymbol{V}$ can be understood as a collection of random variables V satisfying $V(\omega) \in \boldsymbol{V}(\omega)$ $P-$a.s. Such random variables are called selections from the random set $\boldsymbol{V}$, which is made precise in the following definition.

Definition 3.A.3 (Selections, Conditional Selections)**.** *A random element $V : \Omega \to \mathcal{V}$ is called a (measurable) selection of $\boldsymbol{V}$ if $V(\omega) \in \boldsymbol{V}(\omega)$ for $P-$almost all $\omega \in \Omega$. The family of all measurable selections of a random set $\boldsymbol{V}$ will be denoted by $Sel(\boldsymbol{V})$.*

Although it is suppressed in the notation, the family of selections $Sel(\cdot)$ depends both on the distribution of the random set $\boldsymbol{V}$, and on the underlying probability space. Indeed, two identically distributed random sets on the same probability space may have different families of selections.[40] However, the weak closed

[38] See Aliprantis and Border (2006) Ch. 18
[39] See Molchanov (2017) Theorem 1.3.3, p.59.
[40] See Example 1.4.2 in Molchanov (2017), p. 79.

convex hulls of the family of selections from two random closed sets on the same probability space coincide. In addition, when the underlying probability space is non-atomic, it is not necessary to take convex hulls. See the discussion following Definition 3.3.1 in the main text.

3.A.2 PAC Learnability

As described in the introduction, our definition of learnability is related to the definition of learnability prescribed in Valiant (1984). It will thus be useful to understand the concept of learnability from computational learning theory. We will omit technical details in the pursuit of clarity.

In a supervised learning problem, the researcher is presumed to have an i.i.d. sample $\psi = ((y_i, z_i))_{i=1}^n$ from the true measure $P_{Y,Z}$. The researcher is also assumed to have a class of functions $\mathcal{F}$ in mind, called the hypothesis space. The researcher's objective is to select a function $f : \mathcal{Z} \to \mathcal{Y}$, called a hypothesis (or a classifier or a predictor), from the hypothesis space $\mathcal{F}$ that can accurately predict values in $\mathcal{Y}$ given values in $\mathcal{Z}$. The performance of a given function $f \in \mathcal{F}$ is measured according to a loss function. That is, it is assumed the researcher has some function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that $L(y, f(z))$ measures the loss incurred when a prediction $f(z)$ is made and the true value of the outcome is y. The problem of selecting a good hypothesis f is then translated into the problem of choosing $f \in \mathcal{F}$ to minimize expected loss, or risk. A decision rule in this context is a measurable map $d : \Psi_n \to \mathcal{F}$ that selects a hypothesis from the hypothesis space; in learning theory, this decision rule is called an algorithm.

So far the reader should note a resemblance to decision problems seen in statistics and econometrics. However, important differences between the fields arise when evaluating a given statistical decision rule. In particular, computer scientists are interested in rules that achieve close to the minimum possible risk with high probability in finite samples. To define this rigorously, let $\hat{f} \in \mathcal{F}$ be the hypothesis selected by some decision rule (or algorithm) $d : \Psi_n \to \mathcal{F}$. Since $\hat{f} \in \mathcal{F}$ depends on the observed sample, ex-ante it will be a random variable. Now fix any values $(c, \kappa) \in \mathbb{R}_{++} \times (0, 1)$. Then $\hat{f}$ closely approximates the performance of the optimal decision rule in finite samples if:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n} \left(\left| \inf_{f \in \mathcal{F}} \mathbb{E}[L(y, f(z))] - \mathbb{E}[L(y, \hat{f}(z))] \right| \leq c \right) \geq \kappa, \tag{3.102}$$

for a small value of $c \in \mathbb{R}_+$ and a large value of $\kappa \in (0, 1)$ at sample size n. Here $\mathcal{P}_{Y,Z}$ is the collection of all Borel probability measures on $\mathcal{Y} \times \mathcal{Z}$, and thus the performance of a decision rule is uniform over all possible distributions $P_{Y,Z} \in \mathcal{P}_{Y,Z}$.[41] We can now introduce the notion of (agnostic) PAC learnability initially proposed by Haussler (1992).

Definition 3.A.4 (Agnostic PAC Learnability)**.** *A hypothesis class $\mathcal{F}$ is (agnostic) probably approximately correct (PAC) learnable with respect to the loss function L if there exists a function $\zeta_{\mathcal{F}} : \mathbb{R}_+ \times (0, 1) \to \mathbb{N}$ such that, for any $(c, \kappa) \in \mathbb{R}_{++} \times (0, 1) \to \mathbb{N}$, if $n \geq \zeta_{\mathcal{F}}(c, \kappa)$ then there is some decision procedure $d : \Psi_n \to \mathcal{F}$ such that $\hat{f} := d(\psi)$ satisfies (3.102).*

Remark 3.A.1. *This definition omits an important component of the original definition of PAC learnability found in the paper of Valiant (1984), which also requires that the algorithm (decision rule) can be processed in polynomial time (relative to the length of its input). For some this may be a serious omission, as the requirement that an algorithm can be efficiently processed is seen as a core component of learnability in computational learning theory.*[42]

[41] Note that taking the outer probability is necessary because the sampling uncertainty from the choice of $\hat{f}$ is not resolved by the inner expectation.

[42] This perspective is apparent in Valiant (2013).

In other words, a hypothesis space is (agnostic) PAC learnable if we can guarantee that (3.102) holds for any choice of the pair $(c, \kappa) \in \mathbb{R}_{++} \times (0, 1)$ for large enough n. Here c is called the error tolerance parameter, and κ is called the confidence parameter. The "agnostic" component of the definition refers to the fact that the hypothesis class $\mathcal{F}$ may or may not include the true labelling function $f^* : \mathcal{Z} \to \mathcal{Y}$; indeed, such a "true" labelling function may not even exist.

One major advantage of the PAC framework—relative to other frequentist methods of evaluating decision rules—is its analytical tractability and amenability to analysis via concentration inequalities, and techniques from empirical process theory. Indeed, in the case when the decision rule $d : \Psi_n \to \mathcal{F}$ corresponds to the empirical risk minimization rule, it is well known that PAC learnability is implied by uniform convergence (over both $\mathcal{P}_{Y,Z}$ and $\mathcal{F}$) of the empirical risk to the population risk.[43] In specific learning problems this uniform convergence is equivalent to learnability (see the discussion in Alon et al. (1997) and Shalev-Shwartz et al. (2010)). This means well-developed tools in empirical process theory can be used to establish the learnability of a particular class of functions. Intuitively, whether or not a particular class of functions $\mathcal{F}$ is learnable depends on the "complexity" of the function class. There are various ways to measure the complexity of $\mathcal{F}$, some of which are encountered in the current paper. In general, classes that exhibit less complexity are easier to learn than classes that exhibit more complexity, and if a class of functions is too complex, it may not be learnable.

3.A.3 Comparison of PAC Learnability and Minimax Regret

In this subsection we consider a brief comparison between the PAC criterion and the minimax regret criterion in the framework of Wald (1950). Let us consider a more general loss function than the above, given by $L : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$, where $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$. The expected loss when using the function $f \in \mathcal{F}$ when the true probability measure on $\mathcal{X}$ is P is given by:

$$\int L(X, f)\, dP.$$

Now suppose that a decision $d(\psi) = \hat{f}$ is made on the basis of i.i.d. data $\psi = \{(y_i, z_i)\}_{i=1}^n$. Then the *regret* associated with decision $d(\psi)$ when the true probability measure on $\mathcal{X}$ is P is given by:

$$R_n(d, P) := \mathbb{E}_{P^\otimes}\left[\int L(X, d(\psi))\, dP - \inf_{h \in \mathcal{H}} \int L(X, f)\, dP\right].$$

The minimax regret criterion is:

$$\inf_{d \in \mathcal{D}} \sup_{P \in \mathscr{P}} R_n(d, P),$$

where $\mathcal{D}$ is the collection of possible decision rules and $\mathscr{P}$ is the set of all probability measures on $\mathcal{X}$. The following proposition then follows from two simple applications of Markov's inequality:

Proposition 3.A.1. *Suppose that the environment is as above, and that L is uniformly bounded by some value $\overline{L}$. Then $\mathcal{F}$ is (agnostic) PAC-learnable if and only if there exists a decision rule $d \in \mathcal{D}$ such that:*

$$\sup_{P \in \mathscr{P}} R_n(d, P) \to 0,$$

as $n \to \infty$.

[43] See, for example, Shalev-Shwartz and Ben-David (2014) Lemma 4.2.

This proposition shows that any decision rule obtaining zero asymptotic regret uniformly over $\mathscr{P}$ also satisfies the definition of (agnostic) PAC learnability. Furthermore, any decision rule that satisfies the definition of (agnostic) PAC learnability obtains an asymptotic regret of zero.

While this comparison is basic, it demonstrates that there are likely even deeper connections between the PAC learning model of Valiant (1984) and the framework for making statistical decisions proposed by Wald (1950). We believe this to be an important avenue of future research.

Appendix 3.B Proofs

Remark 3.B.1 (Common Notation)**.** *To avoid repetition we introduce some common notation for use in the proofs of Theorem 3.4.1, Theorem 3.5.1, Lemma 3.5.1 and Lemma 3.B.9. In particular, for any $\theta \in \Theta$ and $\gamma \in \Gamma$ let $\lambda^*(\theta,\gamma)$, and $\hat{\lambda}(\theta,\gamma)$ satisfy:*

$$Ph_{\ell b}(\cdot,\theta,\gamma,\lambda^*(\theta,\gamma)) = \max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma,\lambda), \tag{3.103}$$

$$\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\hat{\lambda}(\theta,\gamma)) = \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda). \tag{3.104}$$

Now for any $\gamma \in \Gamma$, let θ^* and $\hat{\theta}$ satisfy:

$$Ph_{\ell b}(\cdot,\theta^*(\gamma),\gamma,\lambda^*(\theta^*(\gamma),\gamma)) \leq \inf_{\theta\in\Theta} Ph_{\ell b}(\cdot,\theta,\gamma,\lambda^*(\theta,\gamma)) + \varepsilon, \tag{3.105}$$

$$\mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\hat{\lambda}(\hat{\theta}(\gamma),\gamma)) \leq \inf_{\theta\in\Theta} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\hat{\lambda}(\theta,\gamma)) + \varepsilon, \tag{3.106}$$

Finally, let γ^* and $\hat{\gamma}$ satisfy:

$$Ph_{\ell b}(\cdot,\theta^*(\gamma^*),\gamma^*,\lambda^*(\theta^*(\gamma^*),\gamma^*)) \geq \sup_{\gamma\in\Gamma} Ph_{\ell b}(\cdot,\theta^*(\gamma),\gamma,\lambda^*(\theta^*(\gamma),\gamma)) - \varepsilon, \tag{3.107}$$

$$\mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\hat{\theta}(\hat{\gamma}),\hat{\gamma})) \geq \sup_{\gamma\in\Gamma} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\hat{\lambda}(\hat{\theta}(\gamma),\gamma)) - \varepsilon. \tag{3.108}$$

With these definitions, it is straightforward to show:

$$\sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma,\lambda) \leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) + 3\varepsilon, \tag{3.109}$$

$$\sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) \leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) + 3\varepsilon. \tag{3.110}$$

Furthermore, we can always choose γ^* and $\hat{\gamma}$ to satisfy:

$$\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda), \tag{3.111}$$

$$\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma^*,\lambda) \leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda). \tag{3.112}$$

Remark 3.B.2 (Measurability)**.** *We will not comment on measurability issues in every proof, and instead we refer readers to the discussion Appendix 3.B.2 (namely, Proposition 3.B.1 and Corollary 3.B.1). There it is shown that certain quantities in this paper that are not typically (Borel) measurable are still universally measurable. This allows us to use outer measures to resolve measurability issues, although this is left implicit in many of the proofs. However, we also note that all measurability issues can also be resolved by restricting Θ and Γ to have at most countably many points.*

3.B.1 Proofs of the Main Results

Proof of Proposition 3.2.1. Recall by assumption we have $\gamma \mapsto \inf_{s\in\mathcal{S}} I[\varphi](\gamma, s)$ is universally measurable. By (Borel) measurability of each decision rule $d : \Psi_n \to \Gamma$ (and thus universal measurability), and the fact that universally measurable functions are closed under composition, this implies that the map $\psi \mapsto \inf_{s\in\mathcal{S}} I[\varphi](d(\psi), s)$ is universally measurable. The result then follows from Lemma 3.B.2 after noting that $\sup_{\gamma\in\Gamma} \inf_{s\in\mathcal{S}} I[\varphi](\gamma, s)$ is a constant for each $P_{Y,Z} \in \mathcal{P}_{Y,Z}$ (and thus plays the role of "$c(P)$" from Lemma 3.B.2). ■

Proof of Lemma 3.3.1. Fix a value of $\delta > 0$ satisfying Assumption 3.3.2. We will focus on proving (3.48) holds, as the proof of (3.49) is similar. By iterated application of Lemma 3.B.3, (3.48) can be rewritten as:

$$\inf_{\theta^*\in\Theta^*} \int \inf_{u\in\boldsymbol{G}^-(y,z,\theta^*)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, dP_{Y,Z} - \int \inf_{u\in\boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, dP_{Y,Z} \leq C_2 d(\theta, \Theta^*).$$

Note that this inequality is trivially satisfied for any $C_2 \geq 0$ when $\theta \in \Theta^*$. Thus, it suffices to focus on the case when $\theta \in \Theta^*_\delta \setminus \Theta^*$. Furthermore, for this latter case it suffices to find a value of $C_2 \geq 0$ satisfying:

$$\int \left(\inf_{u\in\boldsymbol{G}^-(y,z,\theta_1)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_1,\gamma)} \varphi(v) - \inf_{u\in\boldsymbol{G}^-(y,z,\theta_2)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_2,\gamma)} \varphi(v) \right) dP_{Y,Z} \leq C_2 d(\theta_1, \theta_2),$$

for any $\theta_1, \theta_2 \in \Theta^*_\delta$. However, to find C_2 in the previous display, it suffices to find C_2 such that:

$$\inf_{u\in\boldsymbol{G}^-(y,z,\theta_1)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_1,\gamma)} \varphi(v) - \inf_{u\in\boldsymbol{G}^-(y,z,\theta_2)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_2,\gamma)} \varphi(v) \leq C_2 d(\theta_1, \theta_2), \tag{3.113}$$

$(y, z) - a.s.$ Fix any $\varepsilon > 0$ and let $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ be any pair (outside the null sets in (3.50) and (3.51)). For any $\theta_1, \theta_2 \in \Theta^*_\delta$ let u_1^*, u_2^*, y_1^* and y_2^* satisfy:

$$u_1^* \in \boldsymbol{G}^-(y, z, \theta_1), \qquad y_1^* \in \boldsymbol{G}^\star(y, z, u_1^*, \theta_1, \gamma),$$
$$u_2^* \in \boldsymbol{G}^-(y, z, \theta_2), \qquad y_2^* \in \boldsymbol{G}^\star(y, z, u_2^*, \theta_2, \gamma),$$

and:

$$\varphi(y, z, u_1^*, y_1^*) \leq \inf_{u\in\boldsymbol{G}^-(y,z,\theta_1)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_1,\gamma)} \varphi(v) + \varepsilon,$$
$$\varphi(y, z, u_2^*, y_2^*) \leq \inf_{u\in\boldsymbol{G}^-(y,z,\theta_2)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_2,\gamma)} \varphi(v) + \varepsilon.$$

For simplicity we will denote $v_1^* := (y, z, u_1^*, y_1^*)$ and $v_2^* := (y, z, u_2^*, y_2^*)$. Now, by Proposition 3C.1 in Dontchev and Rockafellar (2009), condition (3.51) implies:

$$d_H(\boldsymbol{G}^\star(y, z, u, \theta_1, \gamma), \boldsymbol{G}^\star(y, z, u, \theta_2, \gamma)) \leq \ell_2 d(\theta_1, \theta_2), \qquad \forall \theta_1, \theta_2 \in \Theta^*_\delta$$

$(y, z, u)-$a.s.[44] Thus, since $y_2^* \in \boldsymbol{G}^\star(y, z, u, \theta_2, \gamma)$ by assumption, there exists $y_1 \in \boldsymbol{G}^\star(y, z, u, \theta_1, \gamma)$ such that $d(y_1, y_2^*) \leq \ell_2 d(\theta_1, \theta_2)$. Furthermore, by Proposition 3C.1 in Dontchev and Rockafellar (2009), condition

[44] Recall the Hausdorff distance between two non-empty subsets A and B of a metric space $(\mathcal{X}, d)$ is given by:

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

(3.50) implies:

$$d_H(\boldsymbol{G}^-(y,z,\theta_1),\boldsymbol{G}^-(y,z,\theta_2)) \le \ell_1 d(\theta_1,\theta_2), \qquad \forall \theta_1,\theta_2 \in \Theta^*_\delta.$$

Thus, since $u_2^* \in \boldsymbol{G}^-(y,z,\theta_2)$ by assumption, there exists $u_1 \in \boldsymbol{G}^-(y,z,\theta_1)$ such that $d(u_1,u_2^*) \le \ell_1 d(\theta_1,\theta_2)$. Now let us define $v_1 := (y,z,u_1,y_1)$. Then we have:

$$\begin{aligned}
&\inf_{u\in\boldsymbol{G}^-(y,z,\theta_1)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_1,\gamma)} \varphi(v) - \inf_{u\in\boldsymbol{G}^-(y,z,\theta_2)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta_2,\gamma)} \varphi(v) \\
&\qquad \le \varphi(v_1^*) - \varphi(v_2^*) + \varepsilon \\
&\qquad \le \varphi(v_1) - \varphi(v_2^*) + 2\varepsilon \\
&\qquad \le L_\varphi d((y_1,u_1),(u_2^*,y_2^*)) + 2\varepsilon \\
&\qquad \le L_\varphi \max\{d(y_1,y_2^*), d(u_1,u_2^*)\} + 2\varepsilon \\
&\qquad \le L_\varphi \max\{\ell_1,\ell_2\} d(\theta_1,\theta_2) + 2\varepsilon,
\end{aligned}$$

which holds for all $\theta_1,\theta_2 \in \Theta^*_\delta$.[45] Since $\varepsilon > 0$ is arbitrary, we conclude that C_2 in (3.113) can be taken equal to $L_\varphi \max\{\ell_1,\ell_2\}$. This completes the proof. ■

Proof of Theorem 3.3.1. We will show the lower bound, as the proof for the upper bound is symmetric. We will prove the following sequence of equalities and inequalities:

$$\begin{aligned}
I[\varphi](\gamma) &:= \int \varphi(v)\, dP_{V_\gamma} \\
&\ge \inf_{\theta\in\Theta^*} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} && (3.114) \\
&= \inf_{\theta\in\Theta} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \Bigg(\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} \\
&\qquad\qquad + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta,P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \Bigg) && (3.115) \\
&= \inf_{\theta\in\Theta} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \Bigg(\int \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, dP_{Y,Z,U} \\
&\qquad\qquad + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta,P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \Bigg) && (3.116) \\
&= \inf_{\theta\in\Theta} \int \Bigg(\inf_{u\in\boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v) + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta,P_{Y,Z}) m_j(y,z,u,\theta) \Bigg) dP_{Y,Z} && (3.117) \\
&= \inf_{\theta\in\Theta} \max_{\lambda_j\in\{0,1\}} \int \Bigg(\inf_{u\in\boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v) + \mu^* \sum_{j=1}^J \lambda_j m_j(y,z,u,\theta) \Bigg) dP_{Y,Z}. && (3.118)
\end{aligned}$$

Inequality (3.114) is obvious. Equality (3.115) follows from Lemma 3.B.4. Equalities (3.116) and (3.117) follow from Lemma 3.B.3. Finally, (3.118) follows from Lemma 3.B.5. ■

Proof of Theorem 3.4.1. Let $\mathcal{F}$ be a class of real-valued functions, and let $\psi = ((y_i,z_i))_{i=1}^n$ denote a particular

[45] Here we take the product metric as the sup metric; that is, if $(\mathcal{X},d)$ and $(\mathcal{X}',d')$ are two metric spaces, then the product metric d_∞ on $\mathcal{X}\times\mathcal{X}'$ is defined as $d_\infty((x_1,x_1'),(x_2,x_2')) = \max\{d(x_1,x_2),d'(x_1',x_2')\}$.

sample vector taking values in the sample space Ψ_n. For any $f, f' \in \mathcal{F}$ define the norm:

$$||f - f'||_{\psi,2} := \left(\sum_{i=1}^{n} \left(f(y_i, z_i) - f'(y_i, z_i) \right)^2 \right)^{1/2}.$$

Recall that:

$$h_{\ell b}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right).$$

For notational simplicity we will define:

$$\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) := \frac{1}{n} \sum_{i=1}^{n} \inf_{u_i \in \boldsymbol{G}^-(y_i,z_i,\theta)} \inf_{y_i^\star \in \boldsymbol{G}^\star(y_i,z_i,u_i,\theta,\gamma)} \left(\varphi(v_i) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y_i, z_i, u_i, \theta) \right),$$

$$P h_{\ell b}(\cdot, \theta, \gamma, \lambda) := \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right) dP_{Y,Z}.$$

For any decision rule $d : \Psi_n \to \Gamma$ and any $P_{Y,Z} \in \mathcal{P}_{Y,Z}$, we have by Markov's inequality and Theorem 3.3.1:[46]

$$\begin{aligned} P_{Y,Z}^{\otimes n} \left(\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) \geq c \right) &\leq \frac{1}{c} \mathbb{E} \left(\sup_{\gamma \in \Gamma} \inf_{s \in \mathcal{S}} I[\varphi](\gamma, s) - \inf_{s \in \mathcal{S}} I[\varphi](d(\psi), s) \right) \\ &= \frac{1}{c} \mathbb{E} \left(\sup_{\gamma \in \Gamma} I_{\ell b}[\varphi](\gamma) - I_{\ell b}[\varphi](d(\psi)) \right). \end{aligned} \tag{3.119}$$

Now note by symmetrization (e.g. Van Der Vaart and Wellner (1996) Lemma 2.3.1) we have:

$$\begin{aligned} &\sup_{\gamma \in \Gamma} \sup_{\theta \in \Theta} \max_{\lambda \in \Lambda} \left| \mathbb{E} \left(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right) \right| \\ &\qquad \leq \mathbb{E} \sup_{\gamma \in \Gamma} \sup_{\theta \in \Theta} \max_{\lambda \in \Lambda} \left| \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right| \leq 2\mathbb{E} ||\mathfrak{R}_n|| (\mathcal{H}_{\ell b}), \end{aligned} \tag{3.120}$$

where the final outer expectation is a joint expectation that is also taken over the Rademacher random variables. Now let $\lambda^*(\theta, \gamma)$, $\hat{\lambda}(\theta, \gamma)$, $\theta^*(\gamma)$, $\hat{\theta}(\gamma)$, γ^* and $\hat{\gamma}$ be as in Remark 3.B.1, and set $d(\psi) = \hat{\gamma}$. Then we have:

$$\begin{aligned}
&\mathbb{E} I_{\ell b}[\varphi](d(\psi)) \\
&= \mathbb{E} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, d(\psi), \lambda), && \text{(by Theorem 3.3.1)}, \\
&= \mathbb{E} \inf_{\theta \in \Theta} P h_{\ell b}(\cdot, \theta, d(\psi), \lambda^*(\theta, d(\psi))), && \text{(since } \lambda^* \text{ is optimal at } P \text{ for any } (\theta, \gamma)), \\
&= \mathbb{E} P h_{\ell b}(\cdot, \theta^*(d(\psi)), d(\psi), \lambda^*(\theta^*, d(\psi))) - \varepsilon, && \text{(since } \theta^* \text{ is } \varepsilon\text{−optimal at } (P, \lambda^*) \text{ for any } \gamma), \\
&\geq \mathbb{E} P h_{\ell b}(\cdot, \theta^*(d(\psi)), d(\psi), \hat{\lambda}(\theta^*(d(\psi)), d(\psi))) - \varepsilon, && \text{(since } \lambda^* \text{ was optimal at } P \text{ for any } (\theta, \gamma)), \\
&\geq \mathbb{E} \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(d(\psi)), d(\psi), \hat{\lambda}(\theta^*(d(\psi)), d(\psi))) - 2\mathbb{E} ||\mathfrak{R}_n|| (\mathcal{H}_{\ell b}) - \varepsilon, && \text{(by (3.120))}, \\
&\geq \mathbb{E} \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(d(\psi)), d(\psi), \hat{\lambda}(\hat{\theta}(d(\psi)), d(\psi))) - 2\mathbb{E} ||\mathfrak{R}_n|| (\mathcal{H}_{\ell b}) - 2\varepsilon, && \text{(since } \hat{\theta} \text{ is } \varepsilon\text{-optimal at } (\mathbb{P}_n, \hat{\lambda}) \text{ for any } \gamma), \\
&\geq \mathbb{E} \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \hat{\lambda}(\hat{\theta}(\gamma^*), \gamma^*)) - 2\mathbb{E} ||\mathfrak{R}_n|| (\mathcal{H}_{\ell b}) - 3\varepsilon, && \text{(since } d(\psi) \text{ was } \varepsilon\text{-optimal at } (\mathbb{P}_n, \hat{\lambda}, \hat{\theta})), \\
&\geq \mathbb{E} \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - 2\mathbb{E} ||\mathfrak{R}_n|| (\mathcal{H}_{\ell b}) - 3\varepsilon, && \text{(since } \hat{\lambda} \text{ was optimal at } \mathbb{P}_n \text{ for any } (\theta, \gamma)),
\end{aligned}$$

[46] To be mindful of measurability issues, we can use the outer-measures version of Markov's inequality given in Lemma 6.10 in Kosorok (2008).

$$
\begin{aligned}
&\geq \mathbb{E}Ph_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 3\varepsilon, && \text{(by (3.120)),}\\
&\geq \mathbb{E}Ph_{\ell b}(\cdot, \theta^*(\gamma^*), \gamma^*, \lambda^*(\theta^*(\gamma^*), \gamma^*)) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 4\varepsilon, && \text{(since } \theta^* \text{ is } \varepsilon\text{-optimal at } (P, \lambda^*) \text{ for any } \gamma),\\
&\geq \mathbb{E}\sup_{\gamma\in\Gamma} Ph_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \lambda^*(\theta^*(\gamma), \gamma)) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 5\varepsilon, && \text{(since } \gamma^* \text{ was } \varepsilon\text{-optimal at } (P, \lambda^*, \theta^*)),\\
&\geq \mathbb{E}\sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta} Ph_{\ell b}(\cdot, \theta, \gamma, \lambda^*(\theta, \gamma)) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 5\varepsilon, && \text{(since } \theta^* \text{ was } \varepsilon\text{-optimal at } (P, \lambda^*) \text{ for any } \gamma),\\
&\geq \mathbb{E}\sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 5\varepsilon, && \text{(since } \lambda^* \text{ was optimal at } P \text{ for any } (\theta, \gamma)),\\
&\geq \mathbb{E}\sup_{\gamma\in\Gamma} I_{\ell b}[\varphi](\gamma) - 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) - 5\varepsilon, && \text{(by Theorem 3.3.1).}
\end{aligned}
$$

Since $\varepsilon > 0$ can be taken arbitrarily small, we conclude that:

$$
\mathbb{E}\left(\sup_{\gamma\in\Gamma} I[\varphi](\gamma) - I[\varphi](d(\psi))\right) \leq 4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}). \tag{3.121}
$$

It thus suffices to bound the Rademacher complexity, given by:

$$
\begin{aligned}
&\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b})\\
&= \mathbb{E}\left(\sup_{\gamma\in\Gamma}\sup_{\theta\in\Theta}\max_{\lambda\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\left(\inf_{u_i\in \boldsymbol{G}^-(y_i,z_i,\theta)}\left(\inf_{y_i^\star\in \boldsymbol{G}^\star(y_i,z_i,u_i,\theta,\gamma)}\varphi(v_i) + \mu^*\sum_{j=1}^{J}\lambda_j m_j(y_i,z_i,u_i,\theta)\right)\right)\right|\right).
\end{aligned}
$$

If $\mathcal{H}_{\ell b}$ is not closed under symmetry, then redefine it as $\mathcal{H}_{\ell b} \cup (-\mathcal{H}_{\ell b})$; for our purposes this is without loss of generality, since this operation can only increase the value of $\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b})$. We then have from Lemma 3.B.7 that for any $\varepsilon > 0$:

$$
\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) \leq \frac{2\varepsilon}{\sqrt{n}} + 2\mathrm{Diam}_{\psi,2}(\mathcal{H}_{\ell b})\sqrt{\frac{\log N(\varepsilon, \mathcal{H}_{\ell b}, ||\cdot||_{\psi,2})}{n}}. \tag{3.122}
$$

Since the class of functions $\mathcal{H}_{\ell b}$ is uniformly bounded, we have $\mathrm{Diam}_{\psi,2}(\mathcal{H}_{\ell b}) < \infty$. It remains to bound the metric entropy. To do so, we will define:

$$
\begin{aligned}
\mathcal{H}_I := \Bigg\{ & h(\cdot, u, \theta, \gamma, \lambda) : \mathcal{Y}\times\mathcal{Z} \to \mathbb{R} :\\
& h(y, z, u, \theta, \gamma) = \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v) + \mu^*\sum_{j=1}^{J}\lambda_j m_j(y, z, u, \theta),\\
& \qquad\qquad (u, \theta, \gamma, \lambda) \in \mathcal{U}\times\Theta\times\Gamma\times\Lambda\Bigg\},
\end{aligned} \tag{3.123}
$$

$$
\mathcal{H}_{II} := \left\{h(\cdot, u, \theta, \gamma) : \mathcal{Y}\times\mathcal{Z} \to \mathbb{R} : h(y, z, u, \theta, \gamma) = \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v), \quad (u, \theta, \gamma) \in \mathcal{U}\times\Theta\times\Gamma\right\}, \tag{3.124}
$$

$$
\mathcal{H}_{III} := \left\{h(\cdot, u, y^\star) : \mathcal{Y}\times\mathcal{Z} \to \mathbb{R} : h(y, z, u, y^\star) = \varphi(y, z, u, y^\star), \quad (u, y^\star) \in \mathcal{U}\times\mathcal{Y}^\star\right\}, \tag{3.125}
$$

$$
\mathcal{H}_{IV} := \left\{h(\cdot, u, \theta, \lambda) : \mathcal{Y}\times\mathcal{Z} \to \mathbb{R} : h(y, z, u, \theta) = \sum_{j=1}^{J}\lambda_j m_j(y, z, u, \theta), \ (u, \theta, \lambda) \in \mathcal{U}\times\Theta\times\Lambda\right\}. \tag{3.126}
$$

By Lemma 3.B.6, we have:

$$N(\varepsilon, \mathcal{H}_{\ell b}, ||\cdot||_{\psi,2}) \leq N(\varepsilon/2, \mathcal{H}_I, ||\cdot||_{\psi,2}).$$

By Lemma 3.B.8 we also have:

$$N(\varepsilon/2, \mathcal{H}_I, ||\cdot||_{\psi,2}) \leq N(\varepsilon/2, \mathcal{H}_{II}, ||\cdot||_{\psi,2}) N(\varepsilon/2, \mathcal{H}_{IV}, ||\cdot||_{\psi,2}).$$

Applying Lemma 3.B.6 again we have:

$$N(\varepsilon/2, \mathcal{H}_{II}, ||\cdot||_{\psi,2}) \leq N(\varepsilon/4, \mathcal{H}_{III}, ||\cdot||_{\psi,2}).$$

Finally, from iterated application of Lemma 3.B.8:

$$N(\varepsilon/2, \mathcal{H}_{IV}, ||\cdot||_{\psi,2}) \leq \prod_{j=1}^{J} N(\varepsilon/(2J), \mathcal{M}_j, ||\cdot||_{\psi,2}),$$

We conclude that:

$$\begin{aligned}\log N(\varepsilon, \mathcal{H}_{\ell b}, ||\cdot||_{\psi,2}) &\leq \log N(\varepsilon/4, \mathcal{H}_{III}, ||\cdot||_{\psi,2}) + \sum_{j=1}^{J} \log N(\varepsilon/(2J), \mathcal{M}_j, ||\cdot||_{\psi,2}) \\ &\leq \sup_{Q \in \mathcal{Q}_n} \log N(\varepsilon/4, \mathcal{H}_{III}, ||\cdot||_{Q,2}) + \sum_{j=1}^{J} \sup_{Q \in \mathcal{Q}_n} \log N(\varepsilon/(2J), \mathcal{M}_j, ||\cdot||_{Q,2}),\end{aligned}$$

with the supremum taken over all discrete probability measures $\mathcal{Q}_n$ on $\mathcal{X}$ with atoms that have probabilities that are integer multiples of $1/n$. Since by assumption $\mathcal{H}_{III}$ and $\mathcal{M}_j$ satisfy the entropy growth condition, the right side of the previous display is of order $o(n)$. Combining this with (3.122), we see that for any (c, κ) pair, there exists some n such that $4\mathbb{E}||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) \leq c(1-\kappa)$. Combining this with (3.121) and (3.119), the proof is complete.

■

Proof of Theorem 3.5.1. Recall that:

$$h_{\ell b}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right).$$

For notational simplicity we will define:

$$\begin{aligned}\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) &:= \frac{1}{n} \sum_{i=1}^{n} \inf_{u_i \in \boldsymbol{G}^-(y_i,z_i,\theta)} \inf_{y_i^\star \in \boldsymbol{G}^\star(y_i,z_i,u_i,\theta,\gamma)} \left(\varphi(v_i) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y_i, z_i, u_i, \theta) \right), \\ P h_{\ell b}(\cdot, \theta, \gamma, \lambda) &:= \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right) dP_{Y,Z}.\end{aligned}$$

We claim that it suffices to set $c_n(\kappa) = 2\tilde{c}_n(\psi, \kappa) + 5\varepsilon$, where $\tilde{c}_n(\psi, \kappa)$ satisfies:

$$\sup_{\gamma \in \Gamma} \sup_{\theta \in \Theta} \max_{\lambda \in \Lambda} \left| \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right| \leq \tilde{c}_n(\psi, \kappa), \tag{3.127}$$

with probability at least $\kappa/2$. Let $\lambda^*(\theta, \gamma)$, $\hat{\lambda}(\theta, \gamma)$, $\theta^*(\gamma)$, $\hat{\theta}(\gamma)$, γ^* and $\hat{\gamma}$ be as in Remark 3.B.1 and set

$d(\psi) = \hat{\gamma}$. Then we have:

$$
\begin{aligned}
& I_{\ell b}[\varphi](d(\psi)) \\
&= \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,d(\psi),\lambda), && \text{(by Theorem 3.3.1)},\\
&= \inf_{\theta\in\Theta} Ph_{\ell b}(\cdot,\theta,d(\psi),\lambda^*(\theta,d(\psi))), && \text{(since } \lambda^* \text{ is optimal at } P \text{ for any } (\theta,\gamma)),\\
&= Ph_{\ell b}(\cdot,\theta^*(d(\psi)),d(\psi),\lambda^*(\theta^*,d(\psi))) - \varepsilon, && \text{(since } \theta^* \text{ is } \varepsilon\text{−optimal at } (P,\lambda^*) \text{ for any } \gamma),\\
&\geq Ph_{\ell b}(\cdot,\theta^*(d(\psi)),d(\psi),\hat{\lambda}(\theta^*(d(\psi)),d(\psi))) - \varepsilon, && \text{(since } \lambda^* \text{ was optimal at } P \text{ for any } (\theta,\gamma)),\\
&\geq_{(\kappa/2)} \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(d(\psi)),d(\psi),\hat{\lambda}(\theta^*(d(\psi)),d(\psi))) - \tilde{c}_n(\psi,\kappa) - \varepsilon, && \text{(by (3.127))},\\
&\geq \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(d(\psi)),d(\psi),\hat{\lambda}(\hat{\theta}(d(\psi)),d(\psi))) - \tilde{c}_n(\psi,\kappa) - 2\varepsilon, && \text{(since } \hat{\theta} \text{ is } \varepsilon\text{-optimal at } (\mathbb{P}_n,\hat{\lambda}) \text{ for any } \gamma),\\
&\geq \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\hat{\lambda}(\hat{\theta}(\gamma^*),\gamma^*)) - \tilde{c}_n(\psi,\kappa) - 3\varepsilon, && \text{(since } d(\psi) \text{ was } \varepsilon\text{-optimal at } (\mathbb{P}_n,\hat{\lambda},\hat{\theta})),\\
&\geq \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) - \tilde{c}_n(\psi,\kappa) - 3\varepsilon, && \text{(since } \hat{\lambda} \text{ was optimal at } \mathbb{P}_n \text{ for any } (\theta,\gamma)),\\
&\geq_{(\kappa/2)} Ph_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) - 2\tilde{c}_n(\psi,\kappa) - 3\varepsilon, && \text{(by (3.127))},\\
&\geq Ph_{\ell b}(\cdot,\theta^*(\gamma^*),\gamma^*,\lambda^*(\theta^*(\gamma^*),\gamma^*)) - 2\tilde{c}_n(\psi,\kappa) - 4\varepsilon, && \text{(since } \theta^* \text{ is } \varepsilon\text{-optimal at } (P,\lambda^*) \text{ for any } \gamma),\\
&\geq \sup_{\gamma\in\Gamma} Ph_{\ell b}(\cdot,\theta^*(\gamma),\gamma,\lambda^*(\theta^*(\gamma),\gamma)) - 2\tilde{c}_n(\psi,\kappa) - 5\varepsilon, && \text{(since } \gamma^* \text{ was } \varepsilon\text{-optimal at } (P,\lambda^*,\theta^*)),\\
&\geq \sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta} Ph_{\ell b}(\cdot,\theta,\gamma,\lambda^*(\theta,\gamma)) - 2\tilde{c}_n(\psi,\kappa) - 5\varepsilon, && \text{(since } \theta^* \text{ was } \varepsilon\text{-optimal at } (P,\lambda^*) \text{ for any } \gamma),\\
&\geq \sup_{\gamma\in\Gamma}\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma,\lambda) - 2\tilde{c}_n(\psi,\kappa) - 5\varepsilon, && \text{(since } \lambda^* \text{ was optimal at } P \text{ for any } (\theta,\gamma)),\\
&\geq \sup_{\gamma\in\Gamma} I_{\ell b}[\varphi](\gamma) - 2\tilde{c}_n(\psi,\kappa) - 5\varepsilon, && \text{(by Theorem 3.3.1)}.
\end{aligned}
$$

where each inequality "$\geq_{(\kappa/2)}$" holds with probability at least $\kappa/2$. Note that this shows:

$$\sup_{\gamma\in\Gamma} I_{\ell b}[\varphi](\gamma) - I_{\ell b}[\varphi](\hat{\gamma}) \leq 2\tilde{c}_n(\psi,\kappa) + 5\varepsilon,$$

with probability at least κ. To satisfy (3.127) it clearly suffices to choose $\tilde{c}_n(\psi,\kappa)$ to satisfy:

$$\sup_{P_{Y,Z}\in\mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(\sup_{\gamma\in\Gamma}\sup_{\theta\in\Theta}\max_{\lambda\in\Lambda}\left|\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - Ph_{\ell b}(\cdot,\theta,\gamma,\lambda)\right| \geq \tilde{c}_n(\psi,\kappa)\right) \leq 1-\kappa/2. \tag{3.128}$$

From Koltchinskii (2011) Theorem 4.6 we have for any $t>0$:

$$\sup_{P_{Y,Z}\in\mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(\sup_{\gamma\in\Gamma}\sup_{\theta\in\Theta}\max_{\lambda\in\Lambda}\left|\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - Ph_{\ell b}(\cdot,\theta,\gamma,\lambda)\right| \geq 2||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) + \frac{3t\overline{H}}{\sqrt{n}}\right) \leq \exp\left(-\frac{t^2}{2}\right).$$

Now set:

$$\tilde{c}_n(\psi,\kappa) = 2||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) + \sqrt{\frac{18\ln(2/(2-\kappa))\overline{H}^2}{n}}.$$

Then we have:

$$c_n(\kappa) = 4||\mathfrak{R}_n||(\mathcal{H}_{\ell b}) + \sqrt{\frac{72\ln(2/(2-\kappa))\overline{H}^2}{n}} + 5\varepsilon.$$

Then we conclude (3.83).

■

Proof of Theorem 3.5.2. Let $T, T^\flat$, and $T^\sharp$ be as defined in Lemma 3.5.1. In this proof, it is useful to note the following facts:

(i) The functions $\delta \mapsto T_n(\delta), T(\delta)$ are non-decreasing left-continuous step functions that are greater than or equal to zero on the interval $[0, \delta_0]$, and zero otherwise.

(ii) The functions $\sigma \mapsto T_n^\flat(\sigma), T^\flat(\sigma)$, are non-increasing and left-continuous with their only possible points of discontinuity at the points $\{\delta_j\}_{j=0}^\infty$.

(iii) The functions $\eta \mapsto T_n^\sharp(\eta), T^\sharp(\eta)$ are non-increasing and continuous.

Now for any $\eta > 0$, let:

$$\delta^* = T_n^\sharp(1 - 1/\mathfrak{a}) + \eta',$$
$$\delta^{**} = T^\sharp(1 - 1/\mathfrak{a}) + \eta,$$

where $\eta' = \eta + \varepsilon$ for some $\varepsilon > 0$. Note that choosing δ^{**} slightly larger than $T^\sharp(1 - 1/\mathfrak{a})$ ensures that $T^\flat(\delta^{**}) \leq 1 - 1/\mathfrak{a}$. A similar note applies to δ^* and $T_n^\sharp(1 - 1/\mathfrak{a})$.

From the proof of Lemma 3.5.1 we know there exists an event E_n with $P_{Y,Z}^{\otimes n}(E_n) \geq \kappa$ such that on E_n we have $\mathscr{G}^*(\delta) \subseteq \mathscr{G}_n(\mathfrak{b}\delta)$ for every $\delta \geq \delta^{**}$. Thus, for every $\delta \geq \delta^{**}$ we have on E_n that $T(\delta) \leq T_n(\delta)$, which implies:

$$\frac{T(\delta)}{\delta} \leq \frac{T_n(\delta)}{\delta},$$

for all $\delta \geq \delta^{**}$. Thus, on E_n we have $T^\flat(\sigma) \leq T_n^\flat(\sigma)$ for any $\sigma \geq \delta^{**}$, and in particular we have:

$$T^\flat(\delta^{**}) := \sup_{\delta \geq \delta^{**}} \frac{T(\delta)}{\delta} \leq \sup_{\delta \geq \delta^{**}} \frac{T_n(\delta)}{\delta} =: T_n^\flat(\delta^{**}), \tag{3.129}$$

Recall our choice of δ^{**} ensures that $T^\flat(\delta^{**}) \leq 1 - 1/\mathfrak{a}$. We can now distinguish two cases on the event E_n:

1. We have:

$$\sup_{\delta \geq \delta^{**}} \frac{T(\delta)}{\delta} \leq 1 - \frac{1}{\mathfrak{a}} \leq \sup_{\delta \geq \delta^{**}} \frac{T_n(\delta)}{\delta}.$$

   In this case, we have $T_n^\flat(\delta^{**}) \geq 1 - 1/\mathfrak{a}$, and thus $T_n^\sharp(1 - 1/\mathfrak{a}) \geq \delta^{**}$, so that $\delta^* > \delta^{**}$ (see the definitions of δ^* and δ^{**} above).

2. We have:

$$\sup_{\delta \geq \delta^{**}} \frac{T(\delta)}{\delta} \leq \sup_{\delta \geq \delta^{**}} \frac{T_n(\delta)}{\delta} < 1 - \frac{1}{\mathfrak{a}}.$$

   This implies either (i) $T^\sharp(1 - 1/\mathfrak{a}) \leq T_n^\sharp(1 - 1/\mathfrak{a}) < \delta^{**}$, or (ii) $T_n^\sharp(1 - 1/\mathfrak{a}) < T^\sharp(1 - 1/\mathfrak{a}) < \delta^{**}$. In case (i) we clearly have $\delta^* \geq \delta^{**}$. In case (ii), let:

$$c := T^\sharp(1 - 1/\mathfrak{a}) - T_n^\sharp(1 - 1/\mathfrak{a}) > 0.$$

   Then:

$$\delta^{**} - \delta^* = T^\sharp(1 - 1/\mathfrak{a}) + \eta - T_n^\sharp(1 - 1/\mathfrak{a}) - \eta'$$

$$= c - \varepsilon,$$

where the last line follows from the definition of η'. Now suppose that $c > \varepsilon$ for our $\varepsilon > 0$ chosen at the beginning of the proof. We will show that this produces a contradiction. To understand the approach, note the value of c does not depend on the value of $\eta > 0$, so the assumption that $c > \varepsilon$ must trivially hold for every $\eta > 0$. If we can show that $c < \varepsilon$ for some $\eta > 0$, we will have arrived at our desired contradiction.

Recall that on E_n we have $T^{\flat}(\sigma) \leq T_n^{\flat}(\sigma)$ for any $\sigma \geq \delta^{**}$. This implies that, for any $r > 0$, if $T^{\sharp}(r) \geq \delta^{**}$ then $T_n^{\sharp}(r) \geq T^{\sharp}(r)$. Now choose a value $r_\eta \in \mathbb{R}$ closest to $1 - 1/\mathfrak{a}$ such that $r_\eta \leq 1 - 1/\mathfrak{a}$ and:

$$T^{\sharp}(r_\eta) = T^{\sharp}(1 - 1/\mathfrak{a}) + \eta = \delta^{**}.$$

Such a choice is always possible by continuity of $T^{\sharp}$, and by the fact that $T^{\sharp}$ is non-increasing. By taking η (and thus also δ^{**}) small enough we conclude by continuity of $T^{\sharp}$ that the point r_η can also always be chosen arbitrarily close to $1 - 1/\mathfrak{a}$. Recall by continuity of $T_n^{\sharp}$ that there exists $\varepsilon' > 0$ such that $T_n^{\sharp}(x) - T_n^{\sharp}(1 - 1/\mathfrak{a}) < \varepsilon$ whenever $(1 - 1/\mathfrak{a}) < x + \varepsilon'$. Now by choosing $r_\eta \leq 1 - 1/\mathfrak{a}$ such that $1 - 1/\mathfrak{a} < r_\eta + \varepsilon'$, we have:

$$\begin{aligned} c &= T^{\sharp}(1 - 1/\mathfrak{a}) - T_n^{\sharp}(1 - 1/\mathfrak{a}) \\ &< T^{\sharp}(r_\eta) - T_n^{\sharp}(1 - 1/\mathfrak{a}) \\ &\leq T_n^{\sharp}(r_\eta) - T_n^{\sharp}(1 - 1/\mathfrak{a}) \\ &< \varepsilon. \end{aligned}$$

This of course contradicts the fact that $c > \varepsilon$ for every choice of $\eta > 0$. We conclude that $c \leq \varepsilon$, and since $\delta^{**} - \delta^{*} = c - \varepsilon$, we have $\delta^{**} \leq \delta^{*}$.

We conclude in all cases that $\delta^{**} \leq \delta^{*}$ on E_n. The result then follows directly from Lemma 3.5.1.

■

Proof of Lemma 3.5.1. Recall that:

$$h_{\ell b}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^{-}(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right).$$

For notational simplicity we will define:

$$\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) := \frac{1}{n} \sum_{i=1}^{n} \inf_{u_i \in \boldsymbol{G}^{-}(y_i,z_i,\theta)} \inf_{y_i^\star \in \boldsymbol{G}^\star(y_i,z_i,u_i,\theta,\gamma)} \left(\varphi(v_i) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y_i, z_i, u_i, \theta) \right),$$

$$P h_{\ell b}(\cdot, \theta, \gamma, \lambda) := \int \inf_{u \in \boldsymbol{G}^{-}(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right) dP_{Y,Z}.$$

Define the events:

$$E_{n,j} := \left\{ \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\delta_j)} \sup_{\lambda,\lambda' \in \Lambda} |(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (P h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta', \gamma', \lambda'))| \leq T(\delta_j) \right\},$$

and:

$$E_n := \bigcap_{\{j:\delta_j \geq \delta^{**}\}} E_{n,j}. \tag{3.130}$$

Note the value $2\overline{H}$ is an upper bound for any function in $\mathcal{H}'_{\ell b}(\delta)$ for any $\delta > 0$. By our choice of $\delta_0 > 2\overline{H}$ we have:

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_{n,0}^c\right) = 0.$$

Furthermore, from the uniform version of Hoeffding's inequality (e.g. Koltchinskii (2011) Theorem 4.6, p.71) we have:

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_{n,j}^c\right) \leq \exp\left(-\frac{t_j^2}{2}\right),$$

for each $j \in \mathbb{N}$. We conclude by the union bound that:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_n\right) \geq 1 - \sum_{\{j:\delta_j \geq \delta^{**}\}} \exp\left(-\frac{t_j^2}{2}\right).$$

Now note that with $c_1 = 5$, $c_2 = (3/(2\kappa))^{2/5}$ and $t_j = \sqrt{c_1 \log(c_2 \cdot j)}$, we have:

$$\begin{aligned}
\sum_{\{j:\delta_j \geq \delta^{**}\}} \exp\left(-\frac{t_j^2}{2}\right) &\leq \sum_{j=1}^{\infty} \exp\left(-\frac{t_j^2}{2}\right) \\
&= \sum_{j=1}^{\infty} \exp\left(-\frac{c_1 \log(c_2 \cdot j)}{2}\right) \\
&= \sum_{j=1}^{\infty} (c_2 \cdot j)^{-\frac{c_1}{2}} \\
&= \frac{2(1-\kappa)}{3} \sum_{j=1}^{\infty} \left(\frac{1}{j}\right)^{5/2} \\
&\leq \frac{2(1-\kappa)}{3} \left(\frac{3}{2}\right) \\
&= 1 - \kappa.
\end{aligned}$$

Thus we conclude:

$$\inf_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_n\right) \geq \kappa. \tag{3.131}$$

The remainder of the proof proceeds in two parts:

1. We will show that on the event E_n we have for any $\gamma \in \Gamma$, $\mathscr{E}_n(\gamma) \leq (2 - 1/\mathfrak{a})\left(\mathscr{E}^*(\gamma) \vee \delta^{**}\right)$. We will then use this fact to argue that, on E_n, for any $\delta \geq \delta^{**}$ we have $\mathscr{G}^*(\delta) \subseteq \mathscr{G}_n((2 - 1/\mathfrak{a})\delta)$.
2. We will show that on the event E_n we have for any $\gamma \in \Gamma$, $\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right)$. We will then use this fact to argue that, on E_n, for any $\delta \geq \mathfrak{a}\delta^{**}$ we have $\mathscr{G}_n(\delta/\mathfrak{a}) \subseteq \mathscr{G}^*(\delta)$.

Throughout this proof, let $\lambda^*(\theta, \gamma)$, $\hat{\lambda}(\theta, \gamma)$, $\theta^*(\gamma)$, $\hat{\theta}(\gamma)$, γ^* and $\hat{\gamma}$ be as in Remark 3.B.1.

Part 1: We will prove that on the event E_n we have $\mathscr{E}_n(\gamma) \leq (2 - 1/\mathfrak{a})\,(\mathscr{E}^*(\gamma) \vee \delta^{**})$ for any $\gamma \in \Gamma$. First, consider any γ with $\sigma := \mathscr{E}^*(\gamma) \geq \delta^{**}$. Pick any $\varepsilon > 0$ such that $\delta^{**} \geq \varepsilon$, which is possible since $\delta^{**} > T^\sharp(1 - 1/\mathfrak{a}) \geq 0$. Then on the event E_n we have:

$$
\begin{aligned}
\mathscr{E}_n(\gamma) &:= \sup_{\gamma\in\Gamma} \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) \\
&\leq \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) + 3\varepsilon \\
&= \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) \\
&\qquad + \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) \\
&\qquad\quad - \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) + 3\varepsilon \\
&\leq \sup_{\gamma\in\Gamma} \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) \\
&\qquad + \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) \\
&\qquad\quad - \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) + 3\varepsilon \\
&= \mathscr{E}^*(\gamma) + \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) \\
&\qquad\quad - \left(\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \right) + 3\varepsilon.
\end{aligned}
$$

Now note:

$$
\begin{aligned}
&\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \\
&\leq \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) + \varepsilon \\
&\leq \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda) - \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) + 2\varepsilon \\
&\leq \max_{\lambda\in\Lambda} P h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda) - P h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) + 2\varepsilon \\
&= P h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda^*(\hat{\theta}(\gamma),\gamma)) - P h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) + 2\varepsilon.
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
&\inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) \\
&\leq \inf_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda) + \varepsilon \\
&\leq \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda) + 2\varepsilon \\
&\leq \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) - \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda^*(\hat{\theta}(\gamma),\gamma)) + 2\varepsilon \\
&= \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) - \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda^*(\hat{\theta}(\gamma),\gamma)) + 2\varepsilon.
\end{aligned}
$$

Thus we conclude:

$$
\begin{aligned}
\mathscr{E}_n(\gamma) \leq \mathscr{E}^*(\gamma) + 7\varepsilon &+ P h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda^*(\hat{\theta}(\gamma),\gamma)) - P h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) \\
&- \left(\mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma),\gamma,\lambda^*(\hat{\theta}(\gamma),\gamma)) - \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) \right).
\end{aligned}
$$

However, $\gamma \in \mathscr{G}^*(\sigma)$ by assumption, and by Lemma 3.B.9 we have $\hat{\gamma} \in \mathscr{G}^*(\sigma)$ on the event E_n. Thus, the right side of the previous display can be bounded above:

$$Ph_{\ell b}(\cdot, \hat{\theta}(\gamma), \gamma, \lambda^*(\hat{\theta}(\gamma), \gamma)) - Ph_{\ell b}(\cdot, \theta^*(\hat{\gamma}), \hat{\gamma}, \hat{\lambda}(\theta^*(\hat{\gamma}), \hat{\gamma})) - \left(\mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma), \gamma, \lambda^*(\hat{\theta}(\gamma), \gamma)) - \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(\hat{\gamma})\right)$$
$$\leq \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\sigma)} \sup_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right|.$$

Furthermore, for any $\sigma \geq \delta^{**}$, on the event E_n this final quantity is bounded above by $T(\sigma)$; this follows from the definition of $T(\sigma)$ and the monotonicity of the map:

$$x \mapsto \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(x)} \max_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - Ph_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right|.$$

Thus on E_n:

$$\begin{aligned}
\mathscr{E}_n(\gamma) &\leq \mathscr{E}^*(\gamma) + T(\sigma) + 7\varepsilon \\
&= \mathscr{E}^*(\gamma) + \frac{T(\sigma)}{\sigma}\sigma + 7\varepsilon \\
&\leq \mathscr{E}^*(\gamma) + \sup_{\delta \geq \sigma}\left(\frac{T(\delta)}{\delta}\right)\sigma + 7\varepsilon \\
&= \mathscr{E}^*(\gamma) + T^\flat(\sigma)\sigma + 7\varepsilon \\
&= \mathscr{E}^*(\gamma) + T^\flat(\sigma)\mathscr{E}^*(\gamma) + 7\varepsilon.
\end{aligned}$$

Now, since $\sigma \geq \delta^{**} > T^\sharp(1 - 1/\mathfrak{a})$ we have $T^\flat(\sigma) \leq T^\flat(\delta^{**}) \leq 1 - 1/\mathfrak{a}$. Thus, on the event E_n, if γ is such that $\mathscr{E}^*(\gamma) \geq \delta^{**}$, we have:

$$\mathscr{E}_n(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right)\mathscr{E}^*(\gamma) + 7\varepsilon.$$

Since $\varepsilon > 0$ is any value such that $\delta^{**} \geq \varepsilon$, and thus can be made arbitrarily small, we conclude that on the event E_n we have for any γ with $\mathscr{E}^*(\gamma) \geq \delta^{**}$:

$$\mathscr{E}_n(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right)\mathscr{E}^*(\gamma).$$

Now consider the case when $\sigma := \mathscr{E}^*(\gamma) \leq \delta^{**}$. By the same derivation as above we obtain:

$$\mathscr{E}_n(\gamma)$$
$$\leq \mathscr{E}^*(\gamma) + \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\sigma)} \max_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - Ph_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right| + 7\varepsilon.$$

By monotonicity, we have:

$$\sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\sigma)} \max_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - Ph_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right|$$
$$\leq \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\delta^{**})} \max_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - Ph_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right|.$$

Furthermore, on the event E_n we have:

$$\sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\delta^{**})} \max_{\lambda,\lambda' \in \Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - Ph_{\ell b}(\cdot, \theta', \gamma', \lambda'))\right| \leq T(\delta^{**}).$$

Thus, on the event E_n:

$$\begin{aligned}
\mathscr{E}_n(\gamma) &\leq \mathscr{E}^*(\gamma) + T(\delta^{**}) + 7\varepsilon \\
&\leq \mathscr{E}^*(\gamma) + \sup_{\delta \geq \delta^{**}} \left(\frac{T(\delta)}{\delta} \right) \delta^{**} + 7\varepsilon \\
&= \mathscr{E}^*(\gamma) + T^\flat(\delta^{**})\delta^{**} + 7\varepsilon \\
&\leq \mathscr{E}^*(\gamma) + \left(1 - \frac{1}{\mathfrak{a}}\right) \delta^{**} + 7\varepsilon \\
&\leq \delta^{**} + \left(1 - \frac{1}{\mathfrak{a}}\right) \delta^{**} + 7\varepsilon \\
&= \left(2 - \frac{1}{\mathfrak{a}}\right) \delta^{**} + 7\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is any value such that $\delta^{**} \geq \varepsilon$, and thus can be made arbitrarily small, we conclude that on the event E_n we have for any γ:

$$\mathscr{E}_n(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right) \left(\mathscr{E}^*(\gamma) \vee \delta^{**}\right).$$

We will use this result to argue that, on the event E_n, if $\delta \geq \delta^{**}$ then $\mathscr{E}^*(\gamma) \leq \delta \implies \mathscr{E}_n(\gamma) \leq (2 - 1/\mathfrak{a})\delta$. There are two cases:

(i) $\mathscr{E}^*(\gamma) \leq \delta^{**} \leq \delta$, which implies on the event E_n:

$$\mathscr{E}_n(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right) \left(\mathscr{E}^*(\gamma) \vee \delta^{**}\right) = \left(2 - \frac{1}{\mathfrak{a}}\right) \delta^{**} \leq \left(2 - \frac{1}{\mathfrak{a}}\right) \delta.$$

(ii) $\delta^{**} \leq \mathscr{E}^*(\gamma) \leq \delta$, which implies on the event E_n:

$$\mathscr{E}_n(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right) \left(\mathscr{E}^*(\gamma) \vee \delta^{**}\right) = \left(2 - \frac{1}{\mathfrak{a}}\right) \mathscr{E}^*(\gamma) \leq \left(2 - \frac{1}{\mathfrak{a}}\right) \delta.$$

Thus we conclude that for any $\delta \geq \delta^{**}$, on E_n we have that $\mathscr{E}^*(\gamma) \leq \delta \implies \mathscr{E}_n(\gamma) \leq (2 - 1/\mathfrak{a})\delta$. Now recall that we have $\mathscr{E}^*(\gamma) \leq \delta \iff \gamma \in \mathscr{G}^*(\delta)$ and $\mathscr{E}_n(\gamma) \leq (2 - 1/\mathfrak{a})\delta \iff \gamma \in \mathscr{G}_n((2 - 1/\mathfrak{a})\delta)$. Thus, we conclude that for any $\delta \geq \delta^{**}$, on the event E_n:

$$\mathscr{G}^*(\delta) \subseteq \mathscr{G}_n((2 - 1/\mathfrak{a})\delta),$$

as desired.

Part 2: We will prove that on the event E_n we have $\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right)$ for any $\gamma \in \Gamma$. If γ is such that $\mathscr{E}^*(\gamma) \leq \delta^{**}$ then this is trivially true (since $\mathfrak{a} > 1$). Now consider any γ with $\sigma := \mathscr{E}^*(\gamma) \geq \delta^{**}$. Pick any $\varepsilon > 0$ such that $\delta^{**} \geq \varepsilon$, which is possible since $\delta^{**} > T^\sharp(1 - 1/\mathfrak{a}) \geq 0$. Then on the event E_n we have:

$$\begin{aligned}
\mathscr{E}^*(\gamma) &:= \sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \\
&\leq \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) + 3\varepsilon \\
&= \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \\
&\quad + \left(\sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(\sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right) + 3\varepsilon \\
&= \mathscr{E}^*(\gamma) + \left(\inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right) \\
&\quad - \left(\sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) \right) + 3\varepsilon.
\end{aligned}
$$

Now note:

$$
\begin{aligned}
&\inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma, \lambda) \\
&\leq \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) - \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \lambda) + \varepsilon \\
&\leq \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda) - \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \lambda) + 2\varepsilon \\
&\leq \max_{\lambda \in \Lambda} P h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda) - P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) + 2\varepsilon \\
&\leq P h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) + 2\varepsilon.
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
&\inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) \\
&\leq \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma^*, \lambda) + 3\varepsilon \\
&\leq \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda) + 4\varepsilon \\
&\leq \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \lambda) - \max_{\lambda \in \Lambda} \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda) + 5\varepsilon \\
&\leq \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) - \mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) + 5\varepsilon.
\end{aligned}
$$

Thus we conclude:

$$
\begin{aligned}
\mathscr{E}^*(\gamma) \leq \mathscr{E}_n(\gamma) + 10\varepsilon + P h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) \\
- \left(\mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) \right).
\end{aligned}
$$

However, $\gamma \in \mathscr{G}^*(\sigma)$ by assumption, and $\mathscr{E}^*(\gamma^*) \leq \varepsilon \leq \mathscr{E}^*(\gamma) = \sigma$ implies $\gamma^* \in \mathscr{G}^*(\sigma)$. Thus, the right side of the previous display can be bounded above:

$$
\begin{aligned}
&P h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - P h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) \\
&\qquad - \left(\mathbb{P}_n h_{\ell b}(\cdot, \hat{\theta}(\gamma^*), \gamma^*, \lambda^*(\hat{\theta}(\gamma^*), \gamma^*)) - \mathbb{P}_n h_{\ell b}(\cdot, \theta^*(\gamma), \gamma, \hat{\lambda}(\theta^*(\gamma), \gamma)) \right) \\
&\leq \sup_{\theta, \theta' \in \Theta} \sup_{\gamma, \gamma' \in \mathscr{G}^*(\sigma)} \max_{\lambda, \lambda' \in \Lambda} \left| (\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (P h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta', \gamma', \lambda')) \right|.
\end{aligned}
$$

Furthermore, for any $\sigma \geq \delta^{**}$, on the event E_n this final quantity is bounded above by $T(\sigma)$; this follows from the definition of $T(\sigma)$ and the monotonicity of the map:

$$
x \mapsto \sup_{\theta, \theta' \in \Theta} \sup_{\gamma, \gamma' \in \mathscr{G}^*(x)} \max_{\lambda, \lambda' \in \Lambda} \left| (\mathbb{P}_n h_{\ell b}(\cdot, \theta, \gamma, \lambda) - \mathbb{P}_n h_{\ell b}(\cdot, \theta', \gamma', \lambda')) - (P h_{\ell b}(\cdot, \theta, \gamma, \lambda) - P h_{\ell b}(\cdot, \theta', \gamma', \lambda')) \right|.
$$

Thus on E_n:

$$\begin{aligned}
\mathscr{E}^*(\gamma) &\leq \mathscr{E}_n(\gamma) + T(\sigma) + 10\varepsilon \\
&= \mathscr{E}_n(\gamma) + \frac{T(\sigma)}{\sigma}\sigma + 10\varepsilon \\
&\leq \mathscr{E}_n(\gamma) + \sup_{\delta \geq \sigma}\left(\frac{T(\delta)}{\delta}\right)\sigma + 10\varepsilon \\
&= \mathscr{E}_n(\gamma) + T^\flat(\sigma)\sigma + 10\varepsilon \\
&= \mathscr{E}_n(\gamma) + T^\flat(\sigma)\mathscr{E}^*(\gamma) + 10\varepsilon.
\end{aligned}$$

Now, since $\sigma \geq \delta^{**} > T^\sharp(1 - 1/\mathfrak{a})$ we have $T^\flat(\sigma) \leq T^\flat(\delta^{**}) \leq 1 - 1/\mathfrak{a}$. Thus, on the event E_n, if γ is such that $\sigma = \mathscr{E}^*(\gamma) \geq \delta^{**}$, we have:

$$\begin{aligned}
\mathscr{E}^*(\gamma) &\leq \mathscr{E}_n(\gamma) + \left(1 - \frac{1}{\mathfrak{a}}\right)\mathscr{E}^*(\gamma) + 10\varepsilon \\
&\implies \mathscr{E}^*(\gamma) \leq \mathfrak{a}\mathscr{E}_n(\gamma) + 10\mathfrak{a}\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is any value such that $\delta^{**} \geq \varepsilon$, and thus can be made arbitrarily small, we conclude that on the event E_n we have for any γ:

$$\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right).$$

We will use this result to argue that, on the event E_n, if $\delta/\mathfrak{a} \geq \delta^{**}$ then $\mathscr{E}^*(\gamma) \leq \delta$. There are two cases:

(i) $\mathscr{E}_n(\gamma) \leq \delta^{**} \leq \delta/\mathfrak{a}$, which implies on the event E_n:

$$\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right) = \mathfrak{a}\delta^{**} \leq \delta.$$

(ii) $\delta^{**} \leq \mathscr{E}_n(\gamma) \leq \delta/\mathfrak{a}$, which implies on the event E_n:

$$\mathscr{E}^*(\gamma) \leq \mathfrak{a}\left(\mathscr{E}_n(\gamma) \vee \delta^{**}\right) = \mathfrak{a}\mathscr{E}^*(\gamma) \leq \delta.$$

Thus we conclude that for any $\delta/\mathfrak{a} \geq \delta^{**}$, on E_n we have that $\mathscr{E}_n(\gamma) \leq \delta/\mathfrak{a} \implies \mathscr{E}^*(\gamma) \leq \delta$. Now recall that we have $\mathscr{E}_n(\gamma) \leq \delta/\mathfrak{a} \iff \gamma \in \mathscr{G}_n(\delta/\mathfrak{a})$ and $\mathscr{E}^*(\gamma) \leq \delta \iff \gamma \in \mathscr{G}^*(\delta)$. Thus, we conclude that for any $\delta \geq \mathfrak{a}\delta^{**}$, on the event E_n:

$$\mathscr{G}_n(\delta/\mathfrak{a}) \subseteq \mathscr{G}^*(\delta),$$

as desired. This completes the proof.

■

3.B.2 Auxiliary Results and Proofs

On Issues of Measurability

The following discussion mirrors the discussion in Dudley (2010) Section 3.3 and Dudley (2014) Section 5.3. Let $\mathcal{X}$ be a Polish space, and let $\mathfrak{B}(\mathcal{X})$ be the Borel $\sigma-$algebra on $\mathcal{X}$. Then $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is a measurable space. If P is a probability law on $\mathfrak{B}(\mathcal{X})$, then $(\mathcal{X}, \mathfrak{B}(\mathcal{X}), P)$ is a probability space. Now for any $B \subset \mathcal{X}$, we

can define the outer measure P^* on B as:

$$P^*(B) := \inf\{P(C) : B \subset C \text{ and } C \in \mathfrak{B}(\mathcal{X})\}.$$

By Theorem 3.3.1 in Dudley (2010), there always exists $C \in \mathfrak{B}(\mathcal{X})$ such that $P^*(B) = P(C)$, and such a set C is called a measurable cover of B. Now define the collection of null sets for P as:

$$Null(P) := \{A \subset \mathcal{X} : P^*(A) = 0\}.$$

Furthermore, let $\mathfrak{B}_P^*(\mathcal{X})$ denote the smallest $\sigma-$algebra generated by $\mathfrak{B}(\mathcal{X}) \cup Null(P)$. By Proposition 3.3.2 in Dudley (2010), we have:

$$\mathfrak{B}_P^*(\mathcal{X}) := \{B \subset \mathcal{X} : B \Delta C \in Null(P) \text{ for some } C \in \mathfrak{B}(\mathcal{X})\},$$

where $B\Delta C = (B \setminus C) \cup (C \setminus B)$. We can now extend the measure P from $\mathfrak{B}(\mathcal{X})$ to a measure $\overline{P}$ on $\mathfrak{B}_P^*(\mathcal{X})$ as follows: if $B\Delta C \in Null(P)$ and $C \in \mathfrak{B}(\mathcal{X})$, then set $\overline{P}(B) = P(C)$. Proposition 3.3.3 in Dudley (2010) verifies this is a valid extension; that is, $\overline{P}$ is a measure on $\mathfrak{B}_P^*(\mathcal{X})$ and $\overline{P}$ agrees with P for all sets in $\mathfrak{B}(\mathcal{X})$.

However, note that the collection $\mathfrak{B}_P^*(\mathcal{X})$ clearly depends on the probability measure P. Indeed, if Q is another measure on $\mathfrak{B}(\mathcal{X})$, and $\mathfrak{B}_Q^*(\mathcal{X})$ is defined in an analogous manner to $\mathfrak{B}_P^*(\mathcal{X})$, then it is possible for the two collections $\mathfrak{B}_P^*(\mathcal{X})$ and $\mathfrak{B}_Q^*(\mathcal{X})$ to differ because the null sets of P and Q differ. On the other hand, clearly both $\mathfrak{B}_P^*(\mathcal{X})$ and $\mathfrak{B}_Q^*(\mathcal{X})$ must have many elements in common; for example, both collections must contain the Borel sets $\mathfrak{B}(\mathcal{X})$.

A set $B \in \mathfrak{B}_P^*(\mathcal{X})$ is called *measurable for the completion of* P. If for every probability measure P the set B is measurable for the completion of P, then we call B *universally measurable*. We will denote the universally measurable sets as $\mathfrak{B}^*(\mathcal{X})$; it is easily verified that $\mathfrak{B}^*(\mathcal{X})$ is also a $\sigma-$algebra.[47] By definition, for any two probability measures P and Q, both $\mathfrak{B}_P^*(\mathcal{X})$ and $\mathfrak{B}_Q^*(\mathcal{X})$ contain the universally measurable sets. Also note that, in our example, clearly the Borel sets $\mathfrak{B}(\mathcal{X})$ are universally measurable.

A subset $A \subset \mathcal{X}$ of a Polish space $\mathcal{X}$ (with the Borel $\sigma-$field) is called $\mathfrak{B}(\mathcal{X})$-*analytic* if there exists a compact metric space $\mathcal{Y}$ such that A is the projection onto $\mathcal{X}$ of some $B \in \mathfrak{B}(\mathcal{X}) \otimes \mathfrak{B}(\mathcal{Y})$.[48] A function $f : A \to [-\infty, \infty]$ is called lower (or upper) semi-analytic if A is an analytic set and $\{x \in A : f(x) < c\}$ (or $\{x \in A : f(x) \geq c\}$) is an analytic set for every $c \in \mathbb{R}$; that is, if the epigraph (or hypograph) of f is an analytic set. In a Polish space, every analytic set is universally measurable. We refer the reader to Chapter 8 of Cohn (2013) for additional details.

Lemma 3.B.1 (Infimum over Random Sets is Lower Semi-Analytic)**.** *Suppose that Assumptions 3.2.1, 3.2.2 and 3.2.3 hold. Then for any lower semi-analytic function* $f : \mathcal{V} \times \Gamma \times \Theta \times \{0,1\}^J \to \mathbb{R}$*, the function* $f_{lb,1}(y, z, u, \theta, \gamma, \lambda)$ *given by:*

$$f_{lb,1}(y, z, u, \theta, \gamma, \lambda) := \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} f(v, \theta, \gamma, \lambda), \tag{3.132}$$

is lower semi-analytic; that is, $\{(y, z, u, \gamma, \theta, \lambda) : f_{lb,1}(y, z, u, \theta, \gamma, \lambda) < r\}$ *is an analytic set for every* $r \in \mathbb{R}$*, and thus is universally measurable. In addition, the function* $f_{lb,2}(y, z, \theta, \gamma, \lambda)$ *given by:*

$$f_{lb,2}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} f_{lb,1}(y, z, u, \theta, \gamma, \lambda), \tag{3.133}$$

[47] This follows from the fact that an arbitrary intersection of $\sigma-$algebras is a $\sigma-$algebra.

[48] We note that this is one of many equivalent definitions of an analytic set. See Chapter 8 of Cohn (2013). Our definition is from Stinchcombe and White (1992).

is also lower semi-analytic; that is, $\{(y, z, \theta, \gamma, \lambda) : f_{lb,2}(y, z, \theta, \gamma, \lambda) < r\}$ *is an analytic set for every* $r \in \mathbb{R}$*, and thus is universally measurable.*

Remark 3.B.3. *Defining* $f_{ub,1}(y, z, u, \theta, \gamma, \lambda)$ *and* $f_{ub,2}(y, z, u, \theta, \gamma, \lambda)$ *as the analogous functions with the infimum replaced with the supremum, it is possible to show that* $f_{ub,1}(y, z, u, \theta, \gamma, \lambda)$ *and* $f_{ub,2}(y, z, u, \theta, \gamma, \lambda)$ *are upper semi-analytic.*

Proof of Lemma 3.B.1. Recall that under Assumption 3.2.3, the multifunction $\boldsymbol{G}^\star(y, z, u, \theta, \gamma)$ is Effros measurable with respect to the product Borel $\sigma-$algebra $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\mathcal{U}) \otimes \mathfrak{B}(\Gamma)$. By Molchanov (2017) Theorem 1.3.3 this implies that:

$$\text{Graph}(\boldsymbol{G}^\star) \in \mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\mathcal{U}) \otimes \mathfrak{B}(\Theta) \otimes \mathfrak{B}(\Gamma).$$

Thus $\text{Graph}(\boldsymbol{G}^\star)$ is a Borel (and thus also an analytic) set. Now note that $\boldsymbol{G}^\star(y, z, u, \theta, \gamma)$ can be rewritten as:

$$\boldsymbol{G}^\star(y, z, u, \theta, \gamma) := \{y^\star \in \mathcal{Y}^\star : (y, z, u, y^\star, \theta, \gamma) \in \text{Graph}(\boldsymbol{G}^\star)\}.$$

The fact that $f_{lb,1} : \mathcal{V} \times \Gamma \times \Theta \times \{0, 1\}^J \to \mathbb{R}$ is lower semi-analytic then follows directly from the selection Theorem of Shreve and Bertsekas (1978), p. 968.[49] Taking $f_{lb,1}(y, z, u, \theta, \gamma, \lambda)$ as lower semi-analytic, a nearly identical proof shows that $f_{lb,2}(y, z, \theta, \gamma, \lambda)$ is also lower semi-analytic. ■

Proposition 3.B.1. *Suppose the assumptions of Theorem 3.3.1 hold. Then the maps* $\gamma \mapsto I_{\ell b}[\varphi](\gamma), I_{ub}[\varphi](\gamma)$ *are universally measurable.*

Proof. We will focus on the map $\gamma \mapsto I_{\ell b}[\varphi](\gamma)$, as the proof for the upper envelope function is symmetric. By Theorem 3.3.1 we have:

$$I_{\ell b}[\varphi](\gamma) = \inf_{\theta \in \Theta} \max_{\lambda \in \{0,1\}^J} \int h_{\ell b}(y, z, \theta, \gamma, \lambda)\, dP_{Y,Z},$$

where:

$$h_{\ell b}(y, z, \theta, \gamma, \lambda) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta) \right).$$

Suppose that $h_{\ell b}(y, z, \theta, \gamma, \lambda)$ is lower semi-analytic (we will return to this in a moment). Then by proposition 7.46 in Bertsekas and Shreve (1978), the map:

$$(\theta, \gamma, \lambda) \mapsto \int h_{\ell b}(y, z, \theta, \gamma, \lambda)\, dP_{Y,Z}, \tag{3.134}$$

is lower semi-analytic. Furthermore, suppose that $g_1 : \mathbb{R} \to \mathbb{R}$ and $g_2 : \mathbb{R} \to \mathbb{R}$ are lower semi-analytic. The function $g(x) = g_1(x) \vee g_2(x)$ satisfies:

$$g^{-1}((-\infty, r)) = g_1^{-1}((-\infty, r)) \cup g_2^{-1}((-\infty, r)).$$

Since analytic sets are closed under (countable) unions and intersections (Parthasarathy (2005) Theorem 3.1), we have that g is lower semi-analytic whenever g_1 and g_2 are lower semi-analytic. From this we conclude

[49] See also Bertsekas and Shreve (1978) Proposition 7.47, p. 179.

that the function:

$$(\theta, \gamma) \mapsto \max_{\lambda \in \{0,1\}^J} \int h_{\ell b}(y, z, \theta, \gamma, \lambda)\, dP_{Y,Z},$$

is lower semi-analytic, being the pointwise maximum of at most 2^J lower semi-analytic functions of the form (3.134). Finally, by the selection Theorem of Shreve and Bertsekas (1978), p. 968 (see also Bertsekas and Shreve (1978) Proposition 7.47) we have the map:

$$\gamma \mapsto \sup_{\theta \in \Theta} \max_{\lambda \in \{0,1\}^J} \int h_{\ell b}(y, z, \theta, \gamma)\, dP_{Y,Z},$$

is lower semi-analytic, and thus universally measurable. It thus remains only to show that $h_{\ell b}(y, z, \theta, \gamma, \lambda)$ is lower semi-analytic. By Lemma 3.B.1, $h_{\ell b}(y, z, \theta, \gamma)$ will be lower semi-analytic if we can show the function:

$$(v, \theta, \gamma, \lambda) \mapsto \varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j m_j(y, z, u, \theta), \tag{3.135}$$

is lower semi-analytic. Since both $\varphi(v)$ and $\{m_j(y, z, u, \theta)\}_{j=1}^J$ are Borel measurable by assumption, since the composition of Borel measurable functions are Borel measurable we conclude that (3.135) is Borel measurable. The conclusion then follows from the fact that every Borel measurable function is lower semi-analytic. ■

A nearly identical argument shows that, for every fixed sequence $(\xi_1, \ldots, \xi_n) \in \{-1, 1\}^n$, the Rademacher complexity:

$$((y_1, z_1), \ldots, (y_n, z_n)) \mapsto ||\mathfrak{R}||(\mathcal{H}_{\ell b}),$$

is universally measurable. This is stated as a corollary of the previous result for easy reference.

Corollary 3.B.1. *Suppose the assumptions of Theorem 3.3.1 hold, and suppose that the sequence* (Y_1, Z_1), $\ldots$, (Y_n, Z_n) *are the coordinate projections of the product probability space* $((\mathcal{Y} \times \mathcal{Z})^n, (\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}))^{\otimes n}, P_{Y,Z}^{\otimes n})$. *Then the map:*

$$((Y_1, Z_1), \ldots, (Y_n, Z_n)) \mapsto ||\mathfrak{R}||(\mathcal{H}_{\ell b}),$$

is universally measurable; that is, it is measurable for the completion of $P_{Y,Z}^{\otimes n}$ *for any* $P_{Y,Z} \in \mathcal{P}_{Y,Z}$.

Respect for Weak Dominance of the Preference Relation in Definition 3.2.3

Lemma 3.B.2. *Let* $(\Omega, \mathfrak{A})$ *be a measurable space, and let* $X_1, X_2 : \Omega \times \mathcal{T} \to \mathbb{R}$ *be two stochastic processes such that* $X_1(\cdot, t)$ *and* $X_2(\cdot, t)$ *are measurable for each* t, *and* $\omega \mapsto \inf_{t \in \mathcal{T}} X_1(\omega, t), \inf_{t \in \mathcal{T}} X_2(\omega, t)$ *are universally measurable; that is, measurable with respect to the completion of any probability measure on* $(\Omega, \mathfrak{A})$. *Furthermore, suppose that for any probability measure on* $(\Omega, \mathfrak{A})$ *we have* $X_1(\omega, t) \leq X_2(\omega, t)$ *a.s. for every* $t \in \mathcal{T}$, *and let* $c : \mathcal{P} \to \mathbb{R}_{++}$ *be any value depending only on* P, *where* $\mathcal{P}$ *is the set of all probability measure on* $(\Omega, \mathfrak{A})$. *Finally, let* $c_1, c_2 : (0, 1) \times \mathcal{P} \to \mathbb{R}_{++}$ *be the smallest values satisfying:*

$$P\left(\inf_{t \in \mathcal{T}} X_1(\omega, t) + c_1(\kappa, P) \geq c(P)\right) \geq \kappa,$$
$$P\left(\inf_{t \in \mathcal{T}} X_2(\omega, t) + c_2(\kappa, P) \geq c(P)\right) \geq \kappa,$$

for each $\kappa \in (0, 1)$. *Then for every* $P \in \mathcal{P}$ *we have* $c_2(\kappa, P) \leq c_1(\kappa, P)$ *for every* $\kappa \in (0, 1)$.

Proof. Fix any probability measure $P \in \mathcal{P}$. Then by assumption:

$$X_1(\omega, t) \leq X_2(\omega, t) \ a.s. \qquad \forall t \in \mathcal{T}.$$

This implies:

$$\inf_{t \in \mathcal{T}} X_1(\omega, t) \leq X_2(\omega, t) \ a.s. \qquad \forall t \in \mathcal{T},$$

which in turn implies:

$$\inf_{t \in \mathcal{T}} X_1(\omega, t) \leq \inf_{t \in \mathcal{T}} X_2(\omega, t) \ a.s.,$$

and thus:

$$\inf_{t \in \mathcal{T}} X_1(\omega, t) - c(P) \leq \inf_{t \in \mathcal{T}} X_2(\omega, t) - c(P) \ a.s.$$

Let N denote the null set for which this relation is not true (this set may depend on $P \in \mathcal{P}$). Then we have for every $x \in \mathbb{R}$:

$$\left\{\omega : \inf_{t \in \mathcal{T}} X_1(\omega, t) - c(P) > x\right\} \cap N^c \subseteq \left\{\omega : \inf_{t \in \mathcal{T}} X_2(\omega, t) - c(P) > x\right\} \cap N^c,$$

By assumption, these events belong to the universal $\sigma-$algebra generated by $\mathfrak{A}$, and so are measurable with respect to the completion of any $P \in \mathcal{P}$. This implies that for every $x \in \mathbb{R}$:

$$P\left(\omega : \inf_{t \in \mathcal{T}} X_1(\omega, t) - c(P) > x\right) \leq P\left(\omega : \inf_{t \in \mathcal{T}} X_2(\omega, t) - c(P) > x\right).$$

Now taking any $\kappa \in (0, 1)$ and setting $x = -c_1(\kappa, P)$ we have:

$$\kappa \leq P\left(\omega : \inf_{t \in \mathcal{T}} X_1(\omega, t) + c_1(\kappa, P) > c(P)\right) \leq P\left(\omega : \inf_{t \in \mathcal{T}} X_2(\omega, t) + c_1(\kappa, P) > c(P)\right).$$

By definition, this implies $c_2(\kappa, P)$ can be no larger than $c_1(\kappa, P)$; that is, $c_2(\kappa, P) \leq c_1(\kappa, P)$. Since $\kappa \in (0, 1)$ was arbitrary, we conclude that $c_2(\kappa, P) \leq c_1(\kappa, P)$ for every $\kappa \in (0, 1)$. Since $P \in \mathcal{P}$ was also arbitrary we conclude that for every $P \in \mathcal{P}$ we have $c_2(\kappa, P) \leq c_1(\kappa, P)$ for every $\kappa \in (0, 1)$. This completes the proof. ■

Results on Interchanging Integrals and Supremum/Infimum

Lemma 3.B.3 (Interchange of Integral and Supremum/Infimum)**.** *Let $(\mathcal{V}, \mathfrak{B}(\mathcal{V}))$ and $(\mathcal{V}', \mathfrak{B}(\mathcal{V}'))$ be measurable spaces with $\mathcal{V}$ and $\mathcal{V}'$ as Polish spaces. Let $V \in \mathcal{V}$ be any random variable defined on the probability space $(\Omega, \mathfrak{A}, P)$ with (marginal) distribution $P_V = P \circ V^{-1}$. Furthermore, let $\boldsymbol{G} : \mathcal{V} \to \mathcal{V}'$ be an almost surely non-empty Effros-measurable multifunction. Then for any bounded and measurable function $\varphi : \mathcal{V} \times \mathcal{V}' \to \mathbb{R}$, we have:*

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v') \, dP_V = \sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v)) \, dP_V, \tag{3.136}$$

$$\int \inf_{v' \in \boldsymbol{G}(v)} \varphi(v, v') \, dP_V = \inf_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v)) \, dP_V, \tag{3.137}$$

In particular, if:

$$\mathcal{P}_{V'|V} := \{P_{V'|V} : V' \sim P_{V'|V},\ V' : \mathcal{V} \to \mathcal{V}' \text{ is measurable and } P_{V'|V}(V' \in \boldsymbol{G}(V)|V = v) = 1 \text{ a.s.}\},$$

then:

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V = \sup_{P_{V'|V} \in \mathcal{P}_{V'|V}} \int \varphi(v, v')\, d(P_{V'|V} \times P_V), \tag{3.138}$$

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V = \inf_{P_{V'|V} \in \mathcal{P}_{V'|V}} \int \varphi(v, v')\, d(P_{V'|V} \times P_V). \tag{3.139}$$

Proof of Lemma 3.B.3. Since $\boldsymbol{G}$ is Effros measurable, by Theorem 1.3.3 in Molchanov (2017) we have that $\mathrm{gr}(\boldsymbol{G}) \in \mathfrak{B}(\mathcal{V}) \otimes \mathfrak{B}(\mathcal{V}')$, and thus $\mathrm{gr}(\boldsymbol{G})$ is trivially an analytic set. Now define:

$$\mathrm{gr}_v(\boldsymbol{G}) := \{v' \in \mathcal{V}' : (v, v') \in \mathrm{gr}(\boldsymbol{G})\}.$$

Now let:

$$\varphi^*(v) := \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v') = \sup_{v' \in \mathrm{gr}_v(\boldsymbol{G})} \varphi(v, v').$$

Furthermore, define the set:

$$M := \{v \in \Pi_{\mathcal{V}}(\mathrm{gr}(\boldsymbol{G})) : \exists v' \in \mathrm{gr}_v(\boldsymbol{G}) \text{ s.t. } \varphi(v, v') = \varphi^*(v)\}.$$

where $\Pi_{\mathcal{V}} : \mathcal{V} \times \mathcal{V}' \to \mathcal{V}$ is the projection operator. Fix any $\varepsilon > 0$. By the Exact Selection Theorem (Shreve and Bertsekas (1979), p.16) there exists a universally measurable function $\tilde{v}' : \mathcal{V} \to \mathcal{V}'$ such that $(v, \tilde{v}'(v)) \in \mathrm{gr}(\boldsymbol{G})$ for every $v \in \Pi_{\mathcal{V}}(\mathrm{gr}(\boldsymbol{G}))$ and:

$$\varphi(v, \tilde{v}'(v)) \begin{cases} = \varphi^*(v), & \text{if } v \in M, \\ \geq \varphi^*(v) - \varepsilon, & \text{if } v \notin M. \end{cases}$$

This allows us to write:

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V \leq \int \varphi(v, \tilde{v}'(v))\, dP_V + \varepsilon.$$

Since $\tilde{v}'$ is a (universally) measurable selection, clearly we have:

$$\int \varphi(v, \tilde{v}'(v))\, dP_V \leq \sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V(v))\, dP_V$$

It suffices to show:

$$\sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v))\, dP_V \leq \int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V.$$

For any $\varepsilon > 0$, let $V'_\varepsilon \in Sel(\boldsymbol{G})$ satisfy:

$$\sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v))\, dP_V \leq \int \varphi(v, V'_\varepsilon(v))\, dP_V + \varepsilon.$$

Furthermore, let $N := \{v : V_\varepsilon(v) \notin \boldsymbol{G}(v)\}$. Then by definition of $Sel(\boldsymbol{G})$ we have $P(N) = 0$. Thus:

$$\int \varphi(v, V_\varepsilon(v))\, dP_V = \int_{N^c} \varphi(v, V'_\varepsilon(v))\, dP_V \leq \int_{N^c} \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V \leq \int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V.$$

Combining everything we have:

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V \leq \sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V(v))\, dP_V + \varepsilon \leq \int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V + 2\varepsilon$$

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

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V = \sup_{V' \in Sel_{u.m.}(\boldsymbol{G})} \int \varphi(v, V'(v))\, dP_V.$$

Since each $V' \in Sel_{u.m.}(\boldsymbol{G})$ is universally measurable, each V' can be associated with a $\mathfrak{B}(\mathcal{V})-$measurable random variable $\tilde{V}'$ such that $V' = \tilde{V}'$ a.s. Thus we can conclude:

$$\int \sup_{v' \in \boldsymbol{G}(v)} \varphi(v, v')\, dP_V = \sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v))\, dP_V.$$

To show the final claim, note that for any $V' : \mathcal{V} \to \mathcal{V}'$ we have:

$$P_{V'|V}(V' = v'|V = v) = \mathbb{1}\{V'(v) = v'\}.$$

i.e. the conditional distribution of V' is degenerate. Thus for any $V' \in Sel(\boldsymbol{G})$:

$$\begin{aligned}\int \int \varphi(v, v')\, d(P_{V'|V} \times P_V) &= \int \varphi(v, v') \mathbb{1}\{V'(v) = v'\}\, dP_V \\ &= \int \varphi(v, V'(v))\, dP_V.\end{aligned}$$

By definition of $\mathcal{P}_{V'|V}$, we conclude that:

$$\sup_{V' \in Sel(\boldsymbol{G})} \int \varphi(v, V'(v))\, dP_V = \sup_{P_{V'|V} \in \mathcal{P}_{V'|V}} \int \varphi(v, v')\, d(P_{V'|V} \times P_V).$$

■

Results on Error Bounds

In the next Lemma we focus on the lower envelope function, although clearly an analogous result is true for the upper envelope function. For notational simplicity, denote:

$$\varphi^* := \inf_{\theta \in \Theta^*} \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma}. \tag{3.140}$$

We now have the following result:

Lemma 3.B.4 (Equality Between Primal and Penalized Problems)**.** *Suppose the Assumptions of Theorem 3.3.1 hold. Then there exists functions $\lambda_j^{\ell b} : \Theta \times \mathcal{P}_{Y,Z} \to \{0, 1\}$, $j = 1, \ldots, J$, depending only on θ and the*

distribution $P_{Y,Z}$, such that:

$$\varphi^* = \inf_{\theta\in\Theta} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \left(\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \right).$$

Remark 3.B.4. *Recall that $\mathcal{P}_{Y,Z}$ is the set of all Borel probability measures on $\mathcal{Y}\times\mathcal{Z}$.*

Proof of Lemma 3.B.4. First, note that for any functions $\lambda_j^{\ell b}: \Theta\times\mathcal{P}_{Y,Z}\to\{0,1\}$, $j=1,\ldots,J$, we have:

$$\begin{aligned}
\varphi^* &:= \inf_{\theta\in\Theta^*} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} \\
&= \inf_{\theta\in\Theta^*} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \sup_{\lambda\in\mathbb{R}^J_+} \left(\int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \right) \\
&\geq \inf_{\theta\in\Theta^*} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \left(\int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \right) \\
&\geq \inf_{\theta\in\Theta} \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \left(\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}}[m_j(y,z,u,\theta)] \right).
\end{aligned}$$

It thus suffices to show that there exists functions $\lambda_j^{\ell b}: \Theta\times\mathcal{P}_{Y,Z}\to\{0,1\}$ for $j=1,\ldots,J$ satisfying the reverse inequality. This is done constructively. In particular, define:

$$\begin{aligned}
\mathcal{J}^*(\theta, P_{Y,Z}) := \Big\{ j\in\{1,\ldots,J\} : & \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] \\
& = \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+ \Big\}.
\end{aligned}$$

That is, the set $\mathcal{J}^*(\theta, P_{Y,Z})$ returns the indices of the weakly positive (i.e. weakly violated) moment functions that obtain the inner maximum in the problem:

$$\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+.$$

Now set:

$$\lambda_j^{\ell b}(\theta, P_{Y,Z}) := \mathbb{1}\{ j\in\mathcal{J}^*(\theta, P_{Y,Z}) \}. \tag{3.141}$$

To show why this choice works, begin by fixing any $\theta\in\Theta^*_\delta$. By Assumption 3.3.1(ii) we have:

$$C_2 d(\theta,\Theta^*) \geq \varphi^* - \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma}. \tag{3.142}$$

Furthermore, from Assumption 3.3.1(i), since $\theta\in\Theta^*_\delta$ by assumption, we have:

$$\begin{aligned}
C_1 d(\theta,\Theta^*) &= C_1 \min\{\delta, d(\theta,\Theta^*)\} \\
&\leq \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+ \\
&= \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \lambda_j^{\ell b}(\theta, P_{Y,Z}) |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+ \\
&= \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]
\end{aligned} \tag{3.143}$$

$$\leq \sum_{j=1}^{J} \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}} [m_j(y,z,u,\theta)]$$

$$\leq \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}} [m_j(y,z,u,\theta)]. \tag{3.144}$$

Now by construction we have $\mu^* \geq C_2/C_1$. Thus:

$$C_2 d(\theta, \Theta^*) \leq \mu^* C_1 d(\theta, \Theta^*). \tag{3.145}$$

Now use (3.145) to combine (3.142) and (3.144) and rearrange to obtain:

$$\begin{aligned}
\varphi^* &\leq \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} \\
&\qquad + \mu^* \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}} [m_j(y,z,u,\theta)] \\
&\leq \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \left(\inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}} [m_j(y,z,u,\theta)] \right),
\end{aligned}$$

which holds for all $\theta \in \Theta^*_\delta$. To complete the proof, consider any $\theta \in \Theta \setminus \Theta^*_\delta$. Recall from the assumptions of Theorem 3.3.1 that $\varphi : \mathcal{V} \to [\varphi_{\ell b}, \varphi_{ub}] \subset \mathbb{R}$. Then using Assumption 3.3.1 we have:

$$\begin{aligned}
&\inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \left(\inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}} [m_j(y,z,u,\theta)] \right) \\
&\quad \geq \varphi_{\ell b} + \mu^* \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \int \left(\sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right) d(P_{U|Y,Z} \times P_{Y,Z}) \\
&\quad \geq \varphi_{\ell b} + \mu^* \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}} [m_j(y,z,u,\theta)]|_+ \\
&\quad \geq \varphi_{\ell b} + \mu^* C_1 \min\{\delta, d(\theta, \Theta^*)\} \\
&\quad = \varphi_{\ell b} + \mu^* C_1 \delta \\
&\quad \geq \varphi^*,
\end{aligned}$$

where the last line follows since $\mu^* \geq (\varphi_{ub} - \varphi_{\ell b})/(C_1\delta) \geq (\varphi^* - \varphi_{\ell b})/(C_1\delta)$. Since we have shown the inequality holds for every $\theta \in \Theta$, we have:

$$\inf_{\theta \in \Theta} \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \left(\inf_{P_{Y^\star_\gamma|Y,Z,U} \in \mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, dP_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{Y,Z,U}} [m_j(y,z,u,\theta)] \right) \geq \varphi^*.$$

This completes the proof.

■

Lemma 3.B.5. *Suppose the assumptions of Theorem 3.3.1 hold, and define:*

$$h_{\ell b}(y,z,\theta,\gamma) := \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right).$$

where $\lambda_j^{\ell b} : \Theta \times \mathcal{P}_{Y,Z} \to \{0,1\}$, $j = 1, \ldots, J$, *are as from Lemma 3.B.4. Then we have:*

$$\int h_{\ell b}(y,z,\theta,\gamma)\, dP_{Y,Z} \leq \max_{\lambda_j \in \{0,1\}} \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j m_j(y,z,u,\theta) \right) dP_{Y,Z}, \tag{3.146}$$

with equality holding in (3.146) *if* $\theta \in \Theta^*$.

Proof of Lemma 3.B.5. We have:

$$\begin{aligned}
&\int h_{\ell b}(y,z,\theta,\gamma)\, dP_{Y,Z} \\
&:= \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right) dP_{Y,Z} \\
&= \max_{\lambda_j \in \{0,1\} \text{ s.t } \lambda_j = \lambda_j^{\ell b}(\theta, P_{Y,Z})} \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j m_j(y,z,u,\theta) \right) dP_{Y,Z} \\
&\leq \max_{\lambda_j \in \{0,1\}} \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j m_j(y,z,u,\theta) \right) dP_{Y,Z}.
\end{aligned}$$

The first line holds by definition, the second line holds since the $\lambda_j(\theta, P_{Y,Z})$ depends only on θ, and third line holds because the unconstrained maximum is always weakly larger than the constrained maximum.

It remains only to show that the last inequality holds with equality whenever $\theta \in \Theta^*$. To do so it suffices to show that for any $\theta \in \Theta^*$:

$$\int h_{\ell b}(y,z,\theta,\gamma)\, dP_{Y,Z} \geq \max_{\lambda_j \in \{0,1\}} \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j m_j(y,z,u,\theta) \right) dP_{Y,Z}. \tag{3.147}$$

To this end, note that by Lemma 3.B.3 we have:

$$\begin{aligned}
&\int h_{\ell b}(y,z,\theta,\gamma)\, dP_{Y,Z} \\
&= \int \inf_{u \in \boldsymbol{G}^-(y,z,\theta)} \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right) dP_{Y,Z} \\
&= \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right) d(P_{U|Y,Z} \times P_{Y,Z}).
\end{aligned} \tag{3.148}$$

Now since the infimum of the sum is larger than the sum of the infimums, we have:

$$\begin{aligned}
&\inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \left(\varphi(v) + \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta, P_{Y,Z}) m_j(y,z,u,\theta) \right) d(P_{U|Y,Z} \times P_{Y,Z}) \\
&\qquad \geq \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star \in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, d(P_{U|Y,Z} \times P_{Y,Z}) \\
&\qquad\qquad + \inf_{P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)} \mu^* \sum_{j=1}^J \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z} \times P_{Y,Z}}[m_j(y,z,u,\theta)].
\end{aligned} \tag{3.149}$$

Now recall that $\lambda_j^{\ell b}(\theta, P_{Y,Z}) = 1$ if and only if:

$$\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] = \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+.$$

From here we conclude:

$$\begin{aligned}
&\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+ \\
&= \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] \\
&\leq \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]
\end{aligned}$$

Thus:

$$\begin{aligned}
&\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, dP_{Y,Z,U} \\
&\qquad + \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \mu^* \sum_{j=1}^{J} \lambda_j^{\ell b}(\theta, P_{Y,Z}) \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] \\
&\geq \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, dP_{Y,Z,U} \\
&\qquad + \mu^* \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+. \qquad (3.150)
\end{aligned}$$

However, since $\theta \in \Theta^*$ by assumption, we have:

$$\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \max_{j=1,\ldots,J} |\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]|_+ = 0. \qquad (3.151)$$

Thus, combining (3.148), (3.149), (3.150) and (3.151) we can conclude:

$$\int h_{\ell b}(y,z,\theta,\gamma)\, dP_{Y,Z} \geq \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, d(P_{U|Y,Z}\times P_{Y,Z}). \qquad (3.152)$$

Now, applying Lemma 3.B.3 again we have:

$$\begin{aligned}
&\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \int \inf_{y^\star\in \boldsymbol{G}^\star(y,z,u,\theta,\gamma)} \varphi(v)\, d(P_{U|Y,Z}\times P_{Y,Z}) \\
&\qquad = \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y_\gamma^\star|Y,Z,U}\in\mathcal{P}_{Y_\gamma^\star|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, P_{V_\gamma}. \qquad (3.153)
\end{aligned}$$

Now note for $\theta \in \Theta^*$:

$$\begin{aligned}
&\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y_\gamma^\star|Y,Z,U}\in\mathcal{P}_{Y_\gamma^\star|Y,Z,U}(\theta,\gamma)} \int \varphi(v)\, P_{V_\gamma} \\
&= \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y_\gamma^\star|Y,Z,U}\in\mathcal{P}_{Y_\gamma^\star|Y,Z,U}(\theta,\gamma)} \sup_{\lambda_j\in\mathbb{R}_+} \left(\int \varphi(v)\, P_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] \right) \\
&\geq \inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)} \inf_{P_{Y_\gamma^\star|Y,Z,U}\in\mathcal{P}_{Y_\gamma^\star|Y,Z,U}(\theta,\gamma)} \max_{\lambda_j\in\{0,1\}} \left(\int \varphi(v)\, P_{V_\gamma} + \mu^* \sum_{j=1}^{J} \lambda_j \mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)] \right). \qquad (3.154)
\end{aligned}$$

Now by the minimax inequality:

$$\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)}\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)}\max_{\lambda_j\in\{0,1\}}\left(\int\varphi(v)\,P_{V_\gamma}+\mu^*\sum_{j=1}^{J}\lambda_j\mathbb{E}_{P_{U|Y,Z}\times P_{Y,Z}}[m_j(y,z,u,\theta)]\right)$$

$$\geq\max_{\lambda_j\in\{0,1\}}\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)}\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)}\int\left(\varphi(v)+\mu^*\sum_{j=1}^{J}\lambda_j m_j(y,z,u,\theta)\right)dP_{V_\gamma}. \tag{3.155}$$

Finally, by iterated application of Lemma 3.B.3 we have:

$$\max_{\lambda_j\in\{0,1\}}\inf_{P_{U|Y,Z}\in\mathcal{P}_{U|Y,Z}(\theta)}\inf_{P_{Y^\star_\gamma|Y,Z,U}\in\mathcal{P}_{Y^\star_\gamma|Y,Z,U}(\theta,\gamma)}\int\left(\varphi(v)+\mu^*\sum_{j=1}^{J}\lambda_j m_j(y,z,u,\theta)\right)dP_{V_\gamma}$$

$$\geq\max_{\lambda_j\in\{0,1\}}\int\inf_{u\in\boldsymbol{G}^-(y,z,\theta)}\inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)}\left(\varphi(v)+\mu^*\sum_{j=1}^{J}\lambda_j m_j(y,z,u,\theta)\right)dP_{Y,Z}. \tag{3.156}$$

Combining (3.152), (3.153), (3.154), (3.155), and (3.156) we have:

$$\int h_{\ell b}(y,z,\theta,\gamma)\,dP_{Y,Z}\geq\max_{\lambda_j\in\{0,1\}}\int\inf_{u\in\boldsymbol{G}^-(y,z,\theta)}\inf_{y^\star\in\boldsymbol{G}^\star(y,z,u,\theta,\gamma)}\left(\varphi(v)+\mu^*\sum_{j=1}^{J}\lambda_j m_j(y,z,u,\theta)\right)dP_{Y,Z},$$

whenever $\theta\in\Theta^*$. This concludes the proof. ■

Lemmas Supporting Theorem 3.4.1 on Learnability

Lemma 3.B.6. *Suppose that $\mathcal{F}:=\{f_\alpha(\cdot,\theta):\mathcal{X}\to\mathbb{R}:\theta\in\Theta,\alpha\in\mathcal{A}\}$ is a totally bounded parametric class of measurable real-valued functions on the metric space $(\mathcal{X},d)$, where $(\mathcal{A},d_a)$ and (Θ,d_θ) are also metric spaces. Furthermore let $\mathcal{G}$ be a class of real-valued functions with each element $g(\cdot,\theta):\mathcal{X}\to\mathbb{R}$ defined by:*

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

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

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

Proof of Lemma 3.B.6. As a parametric class of functions (parameterized by (α,θ)), the $\varepsilon/2-$cover of $\mathcal{F}$ can be characterized by a collection of points $\{(\alpha_i,\theta_i)\}_{i=1}^n$, where $n=N(\varepsilon/2,\mathcal{F},||\cdot||_{Q,2})$. Denote such a collection by $\mathcal{N}(\mathcal{F})$. We will show that for any $g\in\mathcal{G}$ there exists a pair $(\alpha',\theta')\in\mathcal{N}(\mathcal{F})$ such that:

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

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

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

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

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

Now let α^* be any value satisfying:

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

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

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

This completes the proof.

■

Lemma 3.B.7. *Let $\mathcal{F}$ be a symmetric class of measurable real-valued functions on a Polish space $\mathcal{X}$, and let $\psi = (x_i)_{i=1}^n$ denote an arbitrary vector of n points from $\mathcal{X}$. Then for any $\varepsilon > 0$:*

$$\mathbb{E}||\mathfrak{R}_n||(\mathcal{F}) \leq \frac{2\varepsilon}{\sqrt{n}} + 2Diam_{\psi,2}(\mathcal{F})\sqrt{\frac{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{\psi,2})}{n}}.$$

Proof of Lemma 3.B.7. Note that:

$$n\mathbb{E}||\mathfrak{R}_n||(\mathcal{F}) = n\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right| = \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(x_i)\right|.$$

Now recall that the Rademacher process $\sum_{i=1}^n \xi_i f(x_i)$ is sub-Gaussian with respect to the euclidean distance between the vectors $(f(x_1), \ldots, f(x_n))$ and $(f'(x_1), \ldots, f'(x_n))$ for $f, f' \in \mathcal{F}$. We denote this euclidean distance by $||f - f'||_{\psi,2}$ to emphasize that the vector $\psi = (x_i)_{i=1}^n$ is fixed. Fix an minimal $\varepsilon-$net $\mathcal{F}^* \subset \mathcal{F}$ in the $||\cdot||_{\psi,2}$ norm. There exists at least one function $f' \in \mathcal{F}^*$ such that:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(x_i)\right| \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| + \varepsilon\sqrt{n}.$$

(For example, we can always take f' to be the element in $\mathcal{F}^*$ to be closest to $-f$ in the $||\cdot||_{\psi,2}$ norm, which is an element of $\mathcal{F}$ by symmetry.) Now for any $f \in \mathcal{F}$, let $f^*(f) \in \mathcal{F}^*$ be a function with $||f - f^*(f)||_{\psi,2} \leq \varepsilon$. Then:

$$\begin{aligned}
&\left|\sum_{i=1}^n \xi_i f(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \\
&= \left|\sum_{i=1}^n \xi_i f(x_i) - \sum_{i=1}^n \xi_i f^*(f)(x_i) + \sum_{i=1}^n \xi_i f^*(f)(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \\
&\leq \left|\sum_{i=1}^n \xi_i f(x_i) - \sum_{i=1}^n \xi_i f^*(f)(x_i)\right| + \left|\sum_{i=1}^n \xi_i f^*(f)(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{||f-f^*||_{\psi,2}\leq\varepsilon} \left|\sum_{i=1}^n \xi_i f(x_i) - \sum_{i=1}^n \xi_i f^*(x_i)\right| + \sup_{f^*,f'\in\mathcal{F}^*} \left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \\
&\leq \sup_{||f-f^*||_{\psi,2}\leq\varepsilon} \sum_{i=1}^n |f(x_i) - f^*(x_i)| + \sup_{f^*,f'\in\mathcal{F}^*} \left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \\
&\leq \sup_{||f-f^*||_{\psi,2}\leq\varepsilon} \sqrt{n}||f-f^*||_{\psi,2} + \sup_{f^*,f'\in\mathcal{F}^*} \left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \\
&\leq \sqrt{n}\varepsilon + \sup_{f^*,f'\in\mathcal{F}^*} \left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|.
\end{aligned}
$$

Note we have used the inequality $||f-f'||_{\psi,1} \leq \sqrt{n}||f-f'||_{\psi,2}$, where $||f-f'||_{\psi,1}$ denotes the L^1 distance between f and f' at the points $\psi = (x_i)_{i=1}^n$. Now for any value $a > 0$ we have:

$$
\begin{aligned}
&\exp\left(a\mathbb{E}\max_{f^*,f'\in\mathcal{F}^*}\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|\right) \\
&\leq \mathbb{E}\exp\left(a\max_{f^*,f'\in\mathcal{F}^*}\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|\right) \\
&= \mathbb{E}\max_{f^*,f'\in\mathcal{F}^*}\exp\left(a\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|\right) \\
&\leq \sum_{f,f^*\in\mathcal{F}^*}\mathbb{E}\exp\left(a\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right|\right) \\
&\leq \sum_{f,f^*\in\mathcal{F}^*}\exp\left(a^2\text{Diam}^2_{\psi,2}(\mathcal{F})/2\right) \\
&\leq N^2(\varepsilon,\mathcal{F},||\cdot||_{\psi,2})\exp\left(a^2\text{Diam}^2_{\psi,2}(\mathcal{F})/2\right),
\end{aligned}
$$

where the second-last inequality follows from the fact that the Rademacher process is sub-Gaussian with parameter $\text{Diam}^2_{\psi,2}(\mathcal{F})$.[50] Taking logs and dividing both sides by $a > 0$, we have:

$$
\mathbb{E}\max_{f^*,f'\in\mathcal{F}^*}\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \leq \frac{2\log N(\varepsilon,\mathcal{F},||\cdot||_{\psi,2})}{a} + \frac{a\text{Diam}^2_{\psi,2}(\mathcal{F})}{2}.
$$

Minimizing the upper bound with respect to "a" yields:[51]

$$
\mathbb{E}\max_{f^*,f'\in\mathcal{F}^*}\left|\sum_{i=1}^n \xi_i f^*(x_i) - \sum_{i=1}^n \xi_i f'(x_i)\right| \leq 2\text{Diam}_{\psi,2}(\mathcal{F})\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{\psi,2})}
$$

We conclude that:

$$
n\mathbb{E}||\mathfrak{R}_n||(\mathcal{F}) \leq 2\sqrt{n}\varepsilon + 2\text{Diam}_{\psi,2}(\mathcal{F})\sqrt{\log N(\varepsilon,\mathcal{F},||\cdot||_{\psi,2})}.
$$

■

[50] Recall a stochastic process $(\omega,t) \mapsto X(\omega,t)$ on a metric space (T,d) is sub-Gaussian with respect to the metric d if $\mathbb{E}\exp(\lambda(X_t - X_s)) \leq \exp(\lambda^2 d(t,s)^2/2)$. For example, the Rademacher process is sub-Gaussian with respect to the euclidean metric.

[51] The minimizing value is $a = 2\left(\log N(\varepsilon,\mathcal{F},||\cdot||_{\psi,2})/\text{Diam}^2_{\psi,2}(\mathcal{F})\right)^{1/2}$.

Lemma 3.B.8. *Let $\mathcal{G}$ and $\mathcal{H}$ be two classes of functions and let $\mathcal{F} := \{g + h : g \in \mathcal{G}, h \in \mathcal{H}\}$. Then:*

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

where $||\cdot||$ is any norm.

Remark 3.B.5. *Note that a nearly identical proof of this result can be used to show that:*

$$N(\varepsilon, \mathcal{F}, ||\cdot||) \leq N(\varepsilon \cdot a, \mathcal{G}, ||\cdot||)N(\varepsilon \cdot b, \mathcal{H}, ||\cdot||),$$

where $a, b > 0$ are any values satisfying $a + b = 1$.

Proof of Lemma 3.B.8. Suppose that $N(\varepsilon/2, \mathcal{G}, ||\cdot||) = n$ and $N(\varepsilon/2, \mathcal{H}, ||\cdot||) = m$. It suffices to show $N(\varepsilon, \mathcal{F}, ||\cdot||) \leq nm$. Let $\mathcal{N}(\mathcal{G})$ denote the centres of the balls that obtain the $n-$cover of $\mathcal{G}$ and let $\mathcal{N}(\mathcal{H})$ denote the centres of the balls that obtain the $m-$cover of $\mathcal{H}$. Enumerate the elements of $\mathcal{N}(\mathcal{G})$ as $g_1, \ldots, g_n$ and enumerate the elements of $\mathcal{N}(\mathcal{H})$ as $h_1, \ldots, h_m$. Now define the following collections:

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

for $j = 1, \ldots, n$ and $k = 1, \ldots, m$. Then $\{G_j\}$ forms a $\varepsilon/2-$cover of $\mathcal{G}$ and $\{H_k\}$ forms a $\varepsilon/2-$cover of $\mathcal{H}$. Now for any $g_j \in \mathcal{N}(\mathcal{G})$ and $h_k \in \mathcal{N}(\mathcal{H})$ let $f_{jk} = g_j + h_k$, and define:

$$F_{jk} := \{f : ||f - f_{jk}|| \leq \varepsilon\}.$$

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

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

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

■

A Lemma Supporting Theorem 3.5.2 and Lemma 3.5.1

Lemma 3.B.9. *Let δ^{**} be as in Lemma 3.5.1. If $\delta \geq \delta^{**} \geq \varepsilon > 0$, then:*

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}(\mathscr{E}^*(\hat{\gamma}) \geq \delta) \leq 1 - \kappa.$$

That is $\hat{\gamma} \in \mathscr{G}^*(\delta)$ with high probability when $\delta \geq \delta^{**} \geq \varepsilon > 0$.

Proof. Throughout this proof, let $\lambda^*(\theta, \gamma)$, $\hat{\lambda}(\theta, \gamma)$, $\theta^*(\gamma)$, $\hat{\theta}(\gamma)$, γ^* and $\hat{\gamma}$ be as in Remark 3.B.1. Fix any $\delta > \delta^{**}$ (the case when $\delta = \delta^{**}$ follows from continuity). If $\sigma := \mathscr{E}^*(\hat{\gamma}) \geq \delta \geq \varepsilon > 0$, then:

$$\mathscr{E}^*(\hat{\gamma}) := \sup_{\gamma \in \Gamma} \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} Ph_{\ell b}(\cdot, \theta, \gamma, \lambda) - \inf_{\theta \in \Theta} \max_{\lambda \in \Lambda} Ph_{\ell b}(\cdot, \theta, \hat{\gamma}, \lambda)$$

$$
\begin{aligned}
&\leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) + 3\varepsilon \\
&= \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \\
&\qquad + \left(\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda)\right) \\
&\qquad\qquad - \left(\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda)\right) + 3\varepsilon \\
&\leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \\
&\qquad - \left(\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda)\right) + 4\varepsilon
\end{aligned}
$$

Now note:

$$
\begin{aligned}
&\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) \\
&\leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta,\gamma^*,\lambda) - \max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) + \varepsilon \\
&\leq \max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda) - \max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) + 2\varepsilon \\
&\leq \max_{\lambda\in\Lambda} Ph_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda) - Ph_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) + 2\varepsilon \\
&\leq Ph_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) - Ph_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) + 2\varepsilon.
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
&\inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma^*,\lambda) \\
&\leq \inf_{\theta\in\Theta}\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta,\hat{\gamma},\lambda) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda) + \varepsilon \\
&\leq \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda) + 2\varepsilon \\
&\leq \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\lambda) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) + 2\varepsilon \\
&\leq \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) - \max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) + 2\varepsilon.
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
\mathscr{E}^*(\hat{\gamma}) \leq\ & Ph_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) - Ph_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma})) \\
&- \left(\max_{\lambda\in\Lambda} \mathbb{P}_n h_{\ell b}(\cdot,\hat{\theta}(\gamma^*),\gamma^*,\lambda^*(\hat{\theta}(\gamma^*),\gamma^*)) - \mathbb{P}_n h_{\ell b}(\cdot,\theta^*(\hat{\gamma}),\hat{\gamma},\hat{\lambda}(\theta^*(\hat{\gamma}),\hat{\gamma}))\right) + 8\varepsilon.
\end{aligned}
$$

Furthermore, $\sigma = \mathscr{E}^*(\hat{\gamma}) \geq \mathscr{E}^*(\gamma^*)$ implies that $\hat{\gamma}, \gamma^* \in \mathscr{G}(\sigma)$. Thus:

$$
\mathscr{E}^*(\hat{\gamma}) \leq \sup_{\theta,\theta'\in\Theta}\sup_{\gamma,\gamma'\in\mathscr{G}^*(\sigma)}\max_{\lambda,\lambda'\in\Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \mathbb{P}_n h_{\ell b}(\cdot,\theta',\gamma',\lambda')) - (Ph_{\ell b}(\cdot,\theta,\gamma,\lambda) - Ph_{\ell b}(\cdot,\theta',\gamma',\lambda'))\right|,
$$

which follows since $\varepsilon > 0$ can be made arbitrarily small. Now define:

$$
E_{n,j} := \left\{\sup_{\theta,\theta'\in\Theta}\sup_{\gamma,\gamma'\in\mathscr{G}^*(\sigma)}\max_{\lambda,\lambda'\in\Lambda} \left|(\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \mathbb{P}_n h_{\ell b}(\cdot,\theta',\gamma',\lambda')) - (Ph_{\ell b}(\cdot,\theta,\gamma,\lambda) - Ph_{\ell b}(\cdot,\theta',\gamma',\lambda'))\right| \leq T(\delta_j)\right\}, \tag{3.157}
$$

and:

$$E_n := \bigcap_{\{j:\delta_j \geq \delta^{**}\}} E_{n,j}.$$

Note by our choice of $\delta_0 > 2\overline{H}$ we have:

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_{n,0}^c\right) = 0.$$

Furthermore, from the uniform version of Hoeffding's inequality (e.g. Koltchinskii (2011) Theorem 4.6, p.71) we have:

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_{n,j}^c\right) \leq \exp\left(-\frac{t_j^2}{2}\right),$$

for each $j \in \mathbb{N}$. We conclude by the union bound that:

$$\sup_{P_{Y,Z} \in \mathcal{P}_{Y,Z}} P_{Y,Z}^{\otimes n}\left(E_n^c\right) \leq \sum_{\{j:\delta_j \geq \delta^{**}\}} \exp\left(-\frac{t_j^2}{2}\right) \leq \sum_{j=0}^{\infty} \exp\left(-\frac{t_j^2}{2}\right) \leq 1 - \kappa.$$

Now on the event E_n, for every $\delta \geq \delta^{**}$ we have:

$$\sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\delta)} \max_{\lambda,\lambda' \in \Lambda} \left|\left(\mathbb{P}_n - P\right)\left(h_{\ell b}(\cdot,\theta,\gamma,\lambda) - h_{\ell b}(\cdot,\theta',\gamma',\lambda')\right)\right| \leq T(\delta).$$

Now suppose by way of contradiction that $\{\mathscr{E}^*(\hat{\gamma}) \geq \delta\} \cap E_n \neq \emptyset$. Then on this event we have:

$$\begin{aligned}
\sigma &:= \mathscr{E}^*(\hat{\gamma}) \\
&\leq \sup_{\theta,\theta' \in \Theta} \sup_{\gamma,\gamma' \in \mathscr{G}^*(\sigma)} \max_{\lambda,\lambda' \in \Lambda} \left|\left(\mathbb{P}_n h_{\ell b}(\cdot,\theta,\gamma,\lambda) - \mathbb{P}_n h_{\ell b}(\cdot,\theta',\gamma',\lambda')\right) - \left(P h_{\ell b}(\cdot,\theta,\gamma,\lambda) - P h_{\ell b}(\cdot,\theta',\gamma',\lambda')\right)\right| \\
&\leq T(\sigma).
\end{aligned}$$

However, note that this implies that $\sigma \leq \delta^{**}$ on the event E_n. But since $\sigma \geq \delta > \delta^{**}$ by assumption, we have a contradiction. We conclude that $\{\mathscr{E}^*(\hat{\gamma})\} \geq \delta\} \cap E_n = \emptyset$, or equivalently that $\{\mathscr{E}^*(\hat{\gamma}) \geq \delta\} \subseteq E_n^c$, where the event E_n^c has probability at most $1 - \kappa$.

■

Appendix 3.C Additional Details for the Examples

3.C.1 Example 1: Simultaneous Discrete Choice

Verification of Assumptions 3.2.1, 3.2.2 and 3.2.3

We will now proceed to verify Assumption 3.2.1, 3.2.2 and 3.2.3. First note that Assumption 3.2.1 is trivially satisfied, since the probability space $(\Omega, \mathfrak{A}, P)$ is complete, and both $\mathcal{U}$ and Θ are compact metric spaces with the euclidean norm.

To verify Assumption 3.2.2, note that the multifunction for the factual domain can be rewritten as:

$$\boldsymbol{G}^-(Y,Z,\theta) = \left\{ u \in \mathcal{U} : \begin{array}{l} u_k \in [\pi_k(Z_k, Y_{-k}; \theta), 1], \text{ if } Y_k = 0, \\ u_k \in [-1, \pi_k(Z_k, Y_{-k}; \theta)], \text{ if } Y_k = 1. \end{array} \right\}. \tag{3.158}$$

From here we conclude that, for any $u \in \mathcal{U}$:

$$d(u, \boldsymbol{G}^{-}(Y,Z,\theta)) = \max_k \left(\mathbb{1}\{Y_k = 0\}|\pi_k(Z_k, Y_{-k}; \theta) - u_k|_+ + \mathbb{1}\{Y_k = 1\}|u_k - \pi_k(Z_k, Y_{-k}; \theta)|_+ \right). \tag{3.159}$$

Under our assumptions, this distance is the maximum of K measurable functions, and so is itself measurable. Since $u \in \mathcal{U}$ was arbitrary, by the result of Himmelberg (1975) (see also Theorem 1.3.3 in Molchanov (2017)) this implies that $\boldsymbol{G}^{-}$ is an Effros-measurable multifunction (w.r.t. $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\Theta)$), as desired. It is then easily seen that the conditional distribution of the vector U given (Y, Z) satisfies (3.4) in Assumption 3.2.2 using the multifunction in (3.158) with $\theta = \theta_0$. To complete the verification of Assumption 3.2.2, note that all the moment functions from the moment conditions in (3.9) and (3.10) are bounded in absolute value and Borel measurable (w.r.t. $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\Theta)$).

We now turn to the verification of Assumption 3.2.3. Recall the counterfactual multifunction:

$$\boldsymbol{G}^{\star}(Z, U, \theta, \gamma) := \left\{ y^{\star} \in \mathcal{Y} : y^{\star}_k = \mathbb{1}\{\pi_k\left(\gamma(Z_k, y^{\star}_{-k}); \theta\right) \geq U_k\},\ k = 1, \ldots, K. \right\}. \tag{3.160}$$

Close inspection reveals that:

$$d(y^{\star}, \boldsymbol{G}^{\star}(Z, U, \theta, \gamma)) = \max_k \left| y^{\star}_k - \mathbb{1}\{\pi_k\left(\gamma(Z_k, y^{\star}_{-k}); \theta\right) \geq U_k\} \right|. \tag{3.161}$$

Under our assumptions, this distance is also the maximum of K measurable functions, and so is itself measurable. Since $y^{\star} \in \mathcal{Y}^{\star}$ was arbitrary, by the result of Himmelberg (1975) (see also Theorem 1.3.3 in Molchanov (2017)) this implies that $\boldsymbol{G}^{\star}$ is an Effros-measurable multifunction (w.r.t. $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\mathcal{Z}) \otimes \mathfrak{B}(\Theta) \otimes \mathfrak{B}(\Theta) \otimes \mathfrak{B}(\Gamma)$), as desired. Finally, it is easily seen that the conditional distribution of the vector $Y^{\star}_{\gamma}$ given (Y, Z, U) satisfies (3.6) in Assumption 3.2.3 using the multifunction in (3.12) with $\theta = \theta_0$.

Verification of Assumption 3.3.1

We will first verify Assumption 3.3.1(ii) for some $C_2 \geq 0$ and $\delta > 0$, and then will show that Assumption 3.3.1(i) is also satisfied for our choice of $\delta > 0$.

As was mentioned in the main text, under our current assumptions for this example Assumption 3.3.1(ii) is not satisfied. The issue is illustrated in Figures 3.5 and 3.6, and a case where Assumption 3.3.1(ii) is satisfied is illustrated in Figure 3.7. The issue arises only when for some $k \in \{1, \ldots, K\}$ and some $z \in \mathcal{Z}$ and $y_{-k} \in \mathcal{Y}_{-k}$ we have: (i) the object of interest is $P(Y^{\star}_{\gamma,k} = 1 | Z_k = z', Y_{-k} = y'_{-k})$ or $P(Y^{\star}_{\gamma,k} = 1)$, (ii) the counterfactual cutoff value $\pi_k(\gamma(z, y_{-k}); \theta^*) = 0$ at some $\theta^* \in \partial\Theta^*$, and (iii) if $P(Y_k = 1 | Z_k = z', Y_{-k} = y'_{-k}) \neq 0.5$, where $(z', y'_{-k}) = \gamma(z, y_{-k})$. In this knife-edge case, a very small change in θ^* to some $\theta \notin \Theta^*$ can cause a discontinuous change in $P(Y^{\star}_{\gamma,k} = 1 | Z_k = z', Y_{-k} = y'_{-k})$ or $P(Y^{\star}_{\gamma,k} = 1)$.

To prevent such discontinuities in the value of the policy transform, we can introduce additional assumptions on the degree of smoothness of the distribution of U_k around zero. In particular, instead of the moment conditions in (3.9) and (3.10) we propose imposing the constraints:

$$P\left(U_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right) - 0.5 \leq \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\}, \tag{3.162}$$

$$0.5 - P\left(U_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right) \leq \max\{-L_0 \pi_k(z', y'_{-k}; \theta), 0\}, \tag{3.163}$$

for some $L_0 > 0$, for $k = 1, \ldots, K$, and for all $z, z' \in \mathcal{Z}$ and $y_{-k}, y'_{-k} \in \mathcal{Y}^{K-1}$. These constraints impose a local Lipschitzian constraint on the distribution of U_k around zero. Note that by taking L_0 sufficiently large, these constraints will only be active when $\pi_k(z', y'_{-k}; \theta)$ is close to zero. It is also easily verified that

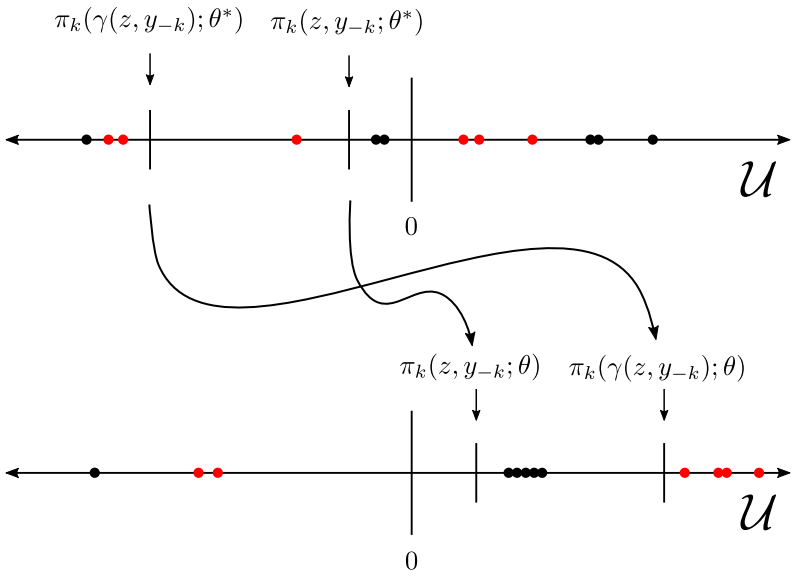


Figure 3.5: This figure illustrates a case violating Assumption 3.3.1(ii). The black dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given (z, y_{-k}). The red dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given $(z', y'_{-k}) = \gamma(z, y_{-k})$. In the upper portion of the figure we have $\theta^* \in \Theta^*$, the median zero assumption is satisfied (three black dots • and three red dots • on either side of zero) and the maximum value of $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k})$ at θ^* is obtained at 1/6. However, in the bottom portion of the figure a small change in the value of $\theta^* \in \Theta^*$ to $\theta \notin \Theta^*$ causes a violation of the median zero assumption for the points (z, y_{-k}) and (z', y'_{-k}). At the new value $\theta \notin \Theta^*$ we have the maximum value of $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k})$ is 1. Note that the scale of the figure can be made arbitrarily small.

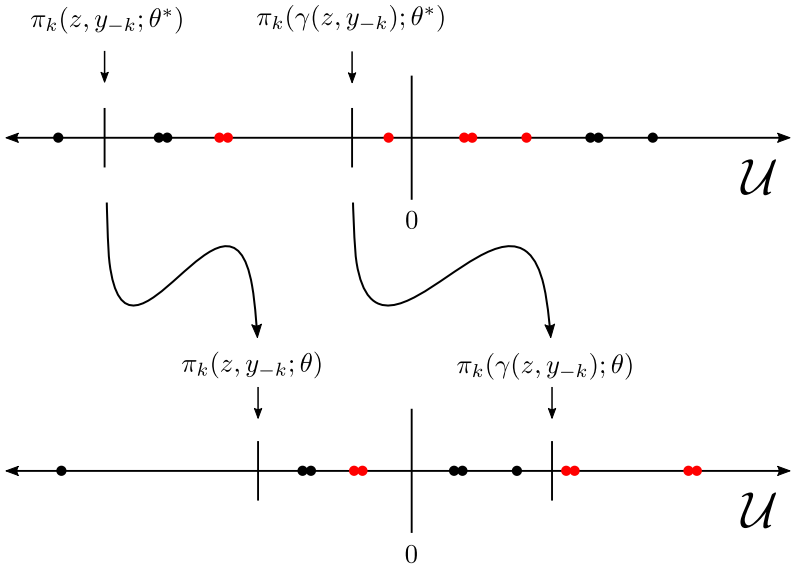


Figure 3.6: This figure illustrates a case violating Assumption 3.3.1(ii). The black dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given (z, y_{-k}). The red dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given $(z', y'_{-k}) = \gamma(z, y_{-k})$. In the upper portion of the figure we have $\theta^* \in \Theta^*$, the median zero assumption is satisfied (three black dots • and three red dots • on either side of zero) and the maximum value of $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k})$ at θ^* is obtained at 1/2. However, in the bottom portion of the figure a small change in the value of $\theta^* \in \Theta^*$ to $\theta \notin \Theta^*$ causes a violation of the median zero assumption for the point (z', y'_{-k}). At the new value $\theta \notin \Theta^*$ we have the maximum value of $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k})$ is 1. Note that the scale of the figure can be made arbitrarily small.

the new moment conditions implied by (3.162) and (3.163) also satisfy Assumption 3.2.2.

We claim that the constraints (3.162) and (3.163) imply that U_k is median zero and median independent

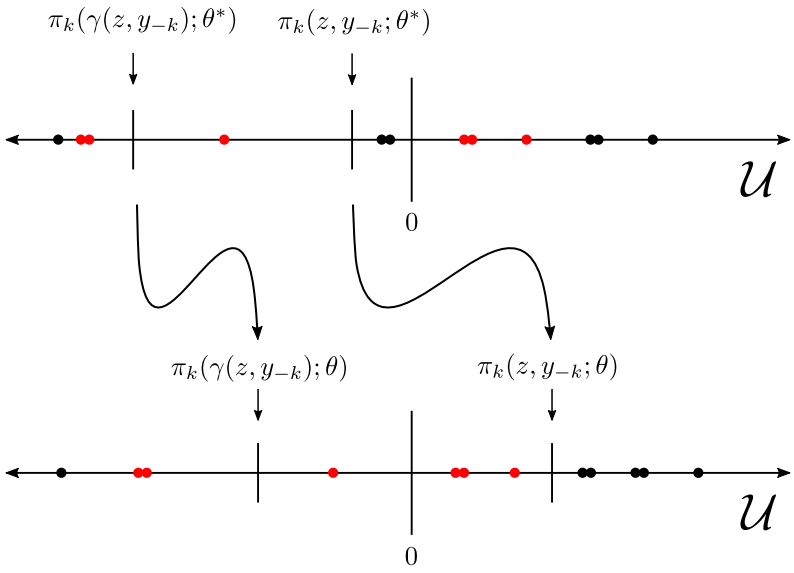


Figure 3.7: This figure illustrates a case that does not violate Assumption 3.3.1(ii). The black dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given (z, y_{-k}). The red dots • represent equal probability masses (1/6) assigned by the conditional distribution of U_k given $(z', y'_{-k}) = \gamma(z, y_{-k})$. In the upper portion of the figure we have $\theta^* \in \Theta^*$, the median zero assumption is satisfied (three black dots • and three red dots • on either side of zero) and $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k}) = 1/6$. In the bottom portion of the figure a small change in the value of $\theta^* \in \Theta^*$ to $\theta \notin \Theta^*$ causes a violation of the median zero assumption for the point (z, y_{-k}). However, at the new value $\theta \notin \Theta^*$ we still have that the maximum obtainable value of $P(Y^\star_\gamma = 1 | Z_k = z, Y_{-k} = y_{-k})$ is 1/6.

of (Z, Y_{-k}). To see this, note that U_k has a median of zero given (z, y_{-k}) if and only if:

(I) $\pi_k(z_k, y_{-k}; \theta) \leq 0$ and $P(U_k \leq \pi_k(z_k, y_{-k}; \theta) | Z = z_k, Y_{-k} = y_{-k}) \leq 0.5$; or

(II) $\pi_k(z_k, y_{-k}; \theta) > 0$ and $P(U_k > \pi_k(z_k, y_{-k}; \theta) | Z = z_k, Y_{-k} = y_{-k}) \leq 0.5$.

The idea behind these conditions is illustrated in Figure 3.8. Conversely, U_k does not have a median of zero conditional on (z, y_{-k}) if and only if:

(i) $\pi_k(z_k, y_{-k}; \theta) > 0$ and $P(U_k \leq \pi_k(z_k, y_{-k}; \theta) | Z = z_k, Y_{-k} = y_{-k}) < 0.5$; or

(ii) $\pi_k(z_k, y_{-k}; \theta) \leq 0$ and $P(U_k > \pi_k(z_k, y_{-k}; \theta) | Z = z_k, Y_{-k} = y_{-k}) < 0.5$.

Note that if (i) holds then (3.163) fails, and if (ii) holds then (3.162) fails. This implies that if both (3.162) and (3.163) hold, then (i) and (ii) do not hold, and thus U_k is median zero and median independent of (Z, Y_{-k}). However, note that it is possible that either (I) or (II) is satisfied but one of (3.162) or (3.163) fails, owing to the fact that together (3.162) and (3.163) are stronger than the median zero and median independence restrictions initially imposed in (3.9) and (3.10).

We will now proceed to verify Assumption 3.3.1. First recall from the discussion in the text that π_k is a known measurable function of (Z_k, Y_{-k}, θ) that is linear in parameters θ and has a gradient (with respect to θ) bounded away from zero for each (z, y_{-k}). Thus, π_k is Lipschitz in θ, and also satisfies a "reverse Lipschitz" condition; that is, for each (z, y_{-k}) we have:

$$L'_k ||\theta - \theta^*|| \leq |\pi_k(z, y_{-k}; \theta) - \pi_k(z, y_{-k}; \theta^*)| \leq L_k ||\theta - \theta^*||,$$

for some $L'_k, L_k > 0$. Now, if one of the constraints (3.162) or (3.163) is violated, we have one of the following inequalities:

$$P(\widetilde{U}_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 > \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\}, \tag{3.164}$$

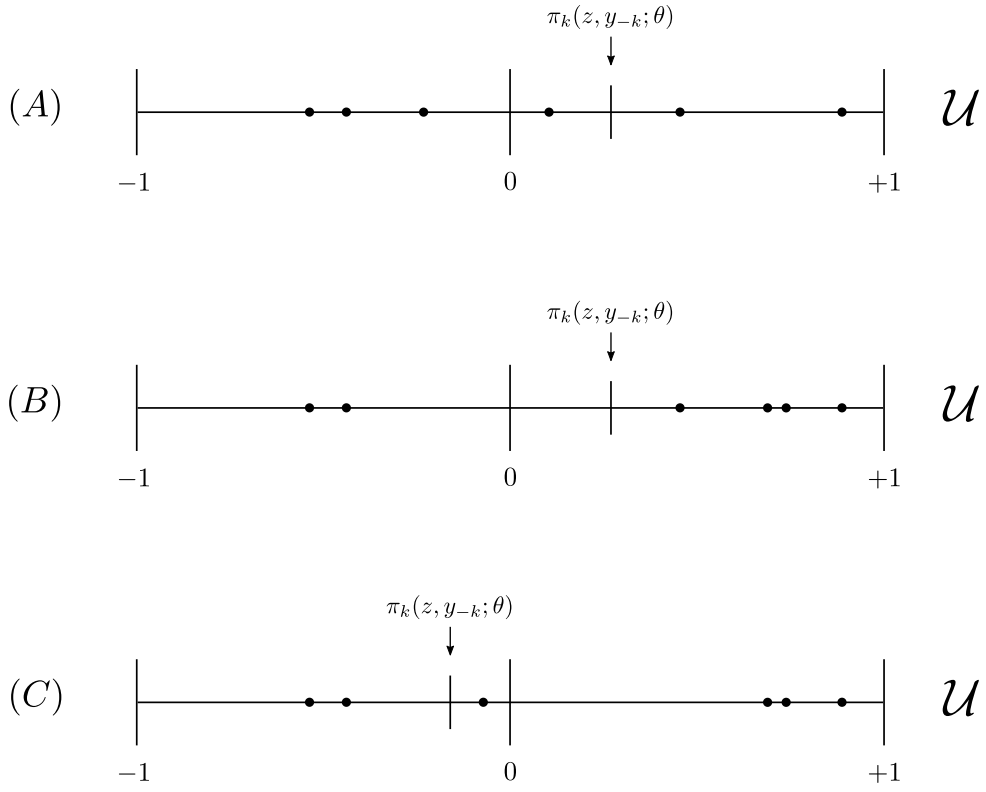


Figure 3.8: This figure illustrates three scenarios, each scenario involving a different allocation of probability mass for U_k, represented by the 6 dots • representing equal probability mass, and a different value of the cutoff $\pi_k(z, y_{-k}; \theta)$. In scenario (A), $\pi_k(z_k, y_{-k}; \theta) > 0$ and $P(Y_k = 0|Z = z_k, Y_{-k} = y_{-k}) \leq 0.5$. In this case the median zero condition can be satisfied, for example, by the allocation of probability mass displayed in the figure. In scenario (B), $\pi_k(z_k, y_{-k}; \theta) > 0$ and $P(Y_k = 0|Z = z_k, Y_{-k} = y_{-k}) > 0.5$. Here there is no way of satisfying the median zero assumption, since too much mass will always be assigned above zero. In scenario (C), $\pi_k(z_k, y_{-k}; \theta) < 0$ and $P(Y_k = 0|Z = z_k, Y_{-k} = y_{-k}) > 0.5$. In this case the median zero condition can again be satisfied, for example, by the allocation of probability mass displayed in the figure.

$$0.5 - P(\widetilde{U}_k \leq \pi_k(z', y'_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}) > \max\{-L_0\pi_k(z', y'_{-k}; \theta), 0\}, \tag{3.165}$$

Subtracting (3.164) from (3.162) and taking $(z', y'_{-k}) = \gamma(z, y_{-k})$, we have:

$$\begin{aligned}
&P\left(U_k \leq \pi_k(\gamma(z, y_{-k}); \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) - P(\widetilde{U}_k \leq \pi_k(\gamma(z, y_{-k}); \theta)|Z_k = z, Y_{-k} = y_{-k}) \\
&\quad = P\left(U_k \leq \pi_k(z', y'_{-k}; \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) - P(\widetilde{U}_k \leq \pi_k(z', y'_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}) \\
&\quad < \max\{L_0\pi_k(z', y'_{-k}; \theta^*), 0\} - \max\{L_0\pi_k(z', y'_{-k}; \theta), 0\} \\
&\quad \leq \max\{L_0\pi_k(z', y'_{-k}; \theta^*) - L_0\pi_k(z', y'_{-k}; \theta), 0\} \\
&\quad \leq L_0|\pi_k(z', y'_{-k}; \theta^*) - \pi_k(z', y'_{-k}; \theta)| \\
&\quad \leq L_0 L_k||\theta - \theta^*||.
\end{aligned} \tag{3.166}$$

Furthermore, subtracting (3.165) from (3.163) and again taking $(z', y'_{-k}) = \gamma(z, y_{-k})$, we have:

$$\begin{aligned}
&P(\widetilde{U}_k \leq \pi_k(\gamma(z, y_{-k}); \theta)|Z_k = z, Y_{-k} = y_{-k}) - P\left(U_k \leq \pi_k(\gamma(z, y_{-k}); \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) \\
&\quad = P(\widetilde{U}_k \leq \pi_k(z', y'_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}) - P\left(U_k \leq \pi_k(z', y'_{-k}; \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) \\
&\quad < \max\{-L_0\pi_k(z', y'_{-k}; \theta^*), 0\} - \max\{-L_0\pi_k(z', y'_{-k}; \theta), 0\} \\
&\quad \leq \max\{L_0\pi_k(z', y'_{-k}; \theta) - L_0\pi_k(z', y'_{-k}; \theta^*), 0\}
\end{aligned}$$

$$\leq L_0|\pi_k(z', y'_{-k}; \theta^*) - \pi_k(z', y'_{-k}; \theta)|$$
$$\leq L_0 L_k ||\theta - \theta^*||. \tag{3.167}$$

From here we can deduce that Assumption 3.3.1(ii) is satisfied for any $\delta > 0$ with $C_2 = L_0 L$ where $L = \min_k L_k$.

To verify Assumption 3.3.1(i), we will first introduce the following Lemma and provide a sketch of its proof:

Lemma 3.C.1. *Consider the simultaneous discrete choice environment of Example 1, but with the new moment conditions (3.162) and (3.163) in place of (3.9) and (3.10). Now fix some value $\theta \in \Theta$. If there exists a random variable U with distribution $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ satisfying:*

$$P\left(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right) - 0.5 \leq \max\{L_0 \pi_k(z, y_{-k}; \theta), 0\}, \tag{3.168}$$
$$0.5 - P\left(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right) \leq \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\}, \tag{3.169}$$

for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$, then $\theta \in \Theta^*$.

Remark 3.C.1. *Note that, precisely because of the result in this Lemma, the new moment conditions implied by (3.162) and (3.163) satisfy the no-backtracking principle from Remark 3.2.1. Indeed, this Lemma shows that (3.168) and (3.169) are sufficient to characterize the identified set. Since these moment conditions do not depend on the counterfactual γ of interest, the no-backtracking principle is satisfied.*

Proof. Note by assumption there exists a random variable U with distribution $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ satisfying (3.168) and (3.169) for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$. Take $\widetilde{U}$ to be a random vector satisfying:

$$P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) = P\left(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right),$$

for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$, so that $\widetilde{U}$ satisfies (3.168) and (3.169). We must show that we can fix probabilities of the form $P(\widetilde{U}_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k})$ for $(z', y'_{-k}) \neq (z, y_{-k})$ in a way that satisfies the remaining constraints from (3.162) and (3.163), as well as the constraints:

$$P(U_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) \leq P(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}),$$

if $\pi_k(z', y'_{-k}; \theta) \leq \pi_k(z, y_{-k}; \theta)$, and:

$$P(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) \leq P(U_k \leq \pi_k(z', y'_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}),$$

if $\pi_k(z, y_{-k}; \theta) \leq \pi_k(z', y'_{-k}; \theta)$. However, such an allocation of probability is clearly always possible. ■

The contrapositive of this result says that if $\theta \notin \Theta^*$, then there is no random variable U with distribution $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ satisfying (3.168) and (3.169); in other words, if $\theta \notin \Theta^*$, then every distribution $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ violates either (3.168) or (3.169). Thus, the Lemma suggests that when analysing violations of the moment conditions in order to verify Assumption 3.3.1(i), it suffices to focus on the moment conditions (3.168) and (3.169).

Finally, there is an important property that will be utilized repeatedly when verifying Assumption 3.3.1(i): for any $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$ and any $P_{U'|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta')$, we must have:

$$P\left(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}\right) = P\left(U'_k \leq \pi_k(z, y_{-k}; \theta') | Z_k = z, Y_{-k} = y_{-k}\right) \tag{3.170}$$

for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$. Indeed, this property follows from the fact that both $P_{U|Y,Z}$ and $P_{U'|Y,Z}$ satisfy the support restrictions for the simultaneous discrete choice model at θ and θ', respectively, and thus they must both rationalize the same observed conditional choice probabilities.

Now we are prepared to verify Assumption 3.3.1(i). First fix some value of $\theta \notin \Theta$. If $\mathcal{P}_{U|Y,Z}(\theta)$ is empty, then Assumption 3.3.1(i) is satisfied for any $C_1, \delta > 0$. Thus, we will focus attention on the non-trivial case where $\mathcal{P}_{U|Y,Z}(\theta)$ is non-empty. Note that if $P(Y_k = 1|Z = z, Y_{-k} = y_{-k}) = 0.5$ for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$, then (3.168) and (3.169) will be satisfied for any $P_{U|Y,Z} \in \mathcal{P}_{U|Y,Z}(\theta)$. By Lemma 3.C.1 this implies $\theta \in \Theta^*$, contradicting the fact that $\theta \notin \Theta$. We conclude that if $P(Y_k = 1|Z = z, Y_{-k} = y_{-k}) = 0.5$ for $k = 1, \ldots, K$ and for every $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$ then $\theta \notin \Theta^*$ implies $\mathcal{P}_{U|Y,Z}(\theta)$ is empty, a case we have ruled out. Thus, we will take as a starting point that there exists at least one k and one pair $(z, y_{-k}) \in \mathcal{Z} \times \mathcal{Y}^{K-1}$ such that $P(Y_k = 1|Z = z, Y_{-k} = y_{-k}) \neq 0.5$. Now define:

$$\tau := \min_k \min_{(z,y_{-k})} |0.5 - P(Y_k = 1|Z = z, Y_{-k} = y_{-k})| \quad s.t. \quad |0.5 - P(Y_k = 1|Z = z, Y_{-k} = y_{-k})| > 0. \tag{3.171}$$

By assumption and by construction we have $\tau > 0$. We now consider violations of the moment conditions (3.168) and (3.169) in turn. First, consider a violation of (3.168). In particular, for our fixed value of $\theta \notin \Theta$ suppose:

$$P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}) - 0.5 > \max\{L_0\pi_k(z, y_{-k}; \theta), 0\}, \tag{3.172}$$

for some k and (z, y_{-k}) pair, where $\widetilde{U}_k$ is a subvector of $\widetilde{U}$ whose distribution is a member of $\mathcal{P}_{U|Y,Z}(\theta)$. Furthermore, let $\theta^* \in \Theta^*$ be the element of Θ^* closest to θ (such an element exists since Θ^* will be closed, which follows from continuity of the payoff functions). There are four cases to consider:

1. $\pi_k(z, y_{-k}; \theta^*) \leq 0$ and $\pi_k(z, y_{-k}; \theta) \leq 0$. Then we have:

$$\max\{L_0\pi_k(z, y_{-k}; \theta), 0\} = 0. \tag{3.173}$$

   However, since $\pi_k(z, y_{-k}; \theta^*) \leq 0$ it must be that:

$$0.5 \geq P\left(U_k \leq \pi_k(z, y_{-k}; \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) = P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}),$$

   where we have used property (3.170) and the fact that θ^* satisfies both (3.162) and (3.163). But then this implies:

$$P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}) - 0.5 \leq 0. \tag{3.174}$$

   Combining (3.173) and (3.174) contradicts the assumption of (3.172). Thus, this case is not possible under the assumption of (3.172).

2. $\pi_k(z, y_{-k}; \theta^*) \leq 0$ and $\pi_k(z, y_{-k}; \theta) > 0$. Then we have:

$$\max\{L_0\pi_k(z, y_{-k}; \theta), 0\} = L_0\pi_k(z, y_{-k}; \theta). \tag{3.175}$$

   However, since $\pi_k(z, y_{-k}; \theta^*) \leq 0$ then it must be that:

$$0.5 \geq P\left(U_k \leq \pi_k(z, y_{-k}; \theta^*)|Z_k = z, Y_{-k} = y_{-k}\right) = P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta)|Z_k = z, Y_{-k} = y_{-k}),$$

where we have used property (3.170) and the fact that θ^* satisfies both (3.162) and (3.163). But then this implies:

$$P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 \leq 0. \tag{3.176}$$

Combining (3.175) and (3.176) contradicts the assumption of (3.172). Thus, this case is not possible under the assumption of (3.172).

3. $\pi_k(z, y_{-k}; \theta^*) > 0$ and $\pi_k(z, y_{-k}; \theta) \leq 0$. Then we have:

$$\max\{L_0 \pi_k(z, y_{-k}; \theta), 0\} = 0.$$

Then:

$$\begin{aligned}
&P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - \max\{L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&= P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 \\
&\geq \tau,
\end{aligned}$$

where the last line follows from the fact that $P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 > 0$ by assumption of (3.172) and the fact $\pi_k(z, y_{-k}; \theta) \leq 0$, and by the definition of τ from (3.57).

4. $\pi_k(z, y_{-k}; \theta^*) > 0$ and $\pi_k(z, y_{-k}; \theta) > 0$. First note that by assumption we have:

$$\begin{aligned}
&P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - \max\{L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&> 0 \\
&\geq P(U_k \leq \pi_k(z, y_{-k}; \theta^*) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - \max\{L_0 \pi_k(z, y_{-k}; \theta^*), 0\}.
\end{aligned}$$

Using (3.170) and the fact that $\pi_k(z, y_{-k}; \theta^*) > 0$ and $\pi_k(z, y_{-k}; \theta) > 0$, this implies $\pi_k(z, y_{-k}; \theta^*) > \pi_k(z, y_{-k}; \theta)$. Now let θ' be a convex combination of θ^* and θ satisfying:

$$P(U'_k \leq \pi_k(z, y_{-k}; \theta') | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - L_0 \pi_k(z, y_{-k}; \theta') = 0,$$

for some selection U'_k. Such an element always exists by linearity of π_k. Then:

$$\begin{aligned}
&P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - \max\{L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&= P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - 0.5 - \max\{L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&\quad + L_0 \pi_k(z, y_{-k}; \theta') - L_0 \pi_k(z, y_{-k}; \theta') \\
&= L_0 \pi_k(z, y_{-k}; \theta') - L_0 \pi_k(z, y_{-k}; \theta) \\
&= L_0 |\pi_k(z, y_{-k}; \theta') - \pi_k(z, y_{-k}; \theta)| \\
&\geq L_0 L'_k ||\theta' - \theta|| \\
&\geq L_0 L'_k ||\theta^* - \theta||.
\end{aligned}$$

In the third last line we used the fact that $\pi_k(z, y_{-k}; \theta^*) > \pi_k(z, y_{-k}; \theta)$. In the second last line we have used the reverse Lipschitz condition, and in the final line we have used the fact that θ' lies between θ and θ^*, by virtue of being a convex combination of these elements.

Next, consider a violation of (3.169). In particular, for our fixed $\theta \notin \Theta$ suppose:

$$0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) > \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\}, \tag{3.177}$$

for some k and (z, y_{-k}) pair, where $\widetilde{U}_k$ is a subvector of $\widetilde{U}$ whose distribution is is a member of $\mathcal{P}_{U|Y,Z}(\theta)$. Again, let $\theta^* \in \Theta^*$ be the element of Θ^* closest to θ. There are again four cases to consider:

1. $\pi_k(z, y_{-k}; \theta^*) \leq 0$ and $\pi_k(z, y_{-k}; \theta) \leq 0$. First note that by assumption we have:

$$\begin{aligned}
&0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&> 0 \\
&\geq 0.5 - P(U_k \leq \pi_k(z, y_{-k}; \theta^*) | Z_k = z, Y_{-k} = y_{-k}) - \max\{-L_0 \pi_k(z, y_{-k}; \theta^*), 0\}.
\end{aligned}$$

   Using (3.170) and the fact that $\pi_k(z, y_{-k}; \theta^*) \leq 0$ and $\pi_k(z, y_{-k}; \theta) \leq 0$, this implies $\pi_k(z, y_{-k}; \theta^*) < \pi_k(z, y_{-k}; \theta)$. Now let θ' be a convex combination of θ^* and θ satisfying:

$$0.5 - P(U'_k \leq \pi_k(z, y_{-k}; \theta') | Z_k = z, Y_{-k} = y_{-k}) + L_0 \pi_k(z, y_{-k}; \theta') = 0,$$

   for some selection U'_k. Such an element always exists by linearity of π_k. Then:

$$\begin{aligned}
&0.5 - P(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&= 0.5 - P(U_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&\quad + L_0 \pi_k(z, y_{-k}; \theta') - L_0 \pi_k(z, y_{-k}; \theta') \\
&= L_0 \pi_k(z, y_{-k}; \theta) - L_0 \pi_k(z, y_{-k}; \theta') \\
&= L_0 |\pi_k(z, y_{-k}; \theta) - \pi_k(z, y_{-k}; \theta')| \\
&\geq L_0 L'_k ||\theta - \theta'|| \\
&\geq L_0 L'_k ||\theta - \theta^*||.
\end{aligned}$$

   In the third last line we used the fact that $\pi_k(z, y_{-k}; \theta^*) < \pi_k(z, y_{-k}; \theta)$. In the second last line we have used the reverse Lipschitz condition, and in the final line we have used the fact that θ' lies between θ and θ^*, by virtue of being a convex combination of these elements.

2. $\pi_k(z, y_{-k}; \theta^*) \leq 0$ and $\pi_k(z, y_{-k}; \theta) > 0$. Then we have:

$$\max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} = 0.$$

   Then:

$$\begin{aligned}
&0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) - \max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} \\
&= 0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) \\
&\geq \tau,
\end{aligned}$$

   where the last line follows from the fact that $0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) > 0$ by assumption of (3.177) and the fact $\pi_k(z, y_{-k}; \theta) > 0$, and by the definition of τ from (3.57).

3. $\pi_k(z, y_{-k}; \theta^*) > 0$ and $\pi_k(z, y_{-k}; \theta) \leq 0$. Then we have:

$$\max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} = -L_0 \pi_k(z, y_{-k}; \theta). \tag{3.178}$$

However, since $\pi_k(z, y_{-k}; \theta^*) > 0$ then it must be that:

$$0.5 \leq P\left(U_k \leq \pi_k(z, y_{-k}; \theta^*) | Z_k = z, Y_{-k} = y_{-k}\right) = P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}),$$

where we have used property (3.170) and the fact that θ^* satisfies both (3.162) and (3.163). But then this implies:

$$0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) \leq 0. \tag{3.179}$$

Combining (3.178) and (3.179) contradicts the assumption of (3.177). Thus, this case is not possible under the assumption of (3.177).

4. $\pi_k(z, y_{-k}; \theta^*) > 0$ and $\pi_k(z, y_{-k}; \theta) > 0$. Then we have:

$$\max\{-L_0 \pi_k(z, y_{-k}; \theta), 0\} = 0. \tag{3.180}$$

However, since $\pi_k(z, y_{-k}; \theta^*) > 0$ then it must be that:

$$0.5 \leq P\left(U_k \leq \pi_k(z, y_{-k}; \theta^*) | Z_k = z, Y_{-k} = y_{-k}\right) = P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}),$$

where we have used property (3.170) and the fact that θ^* satisfies both (3.162) and (3.163). But then this implies:

$$0.5 - P(\widetilde{U}_k \leq \pi_k(z, y_{-k}; \theta) | Z_k = z, Y_{-k} = y_{-k}) \leq 0. \tag{3.181}$$

Combining (3.180) and (3.181) contradicts the assumption of (3.177). Thus, this case is not possible under the assumption of (3.177).

Combining everything, we conclude that Assumption 3.3.1 holds with $C_1 = L_0 L'$ and $\delta = \tau/(L_0 L')$, where $L' = \min_k L'_k$.

Verification of Learnability

By the assumed linearity of π_k with respect to θ, and since π_k depends only on the subvector θ_k of θ, the function $(u, \theta) \mapsto \pi_k(\gamma(z, y_{-k}); \theta) - u$ is a hyperplane in $\mathbb{R}^{d_k}$ for each (z, y_{-k}), where d_k is the dimension of θ_k. By Lemma 2.6.15 in Van Der Vaart and Wellner (1996), for example, Φ is a Vapnik-Chervonenkis (VC) class with VC dimension at most $d_k + 2$. Furthermore, recalle that Φ can be taken to be uniformly bounded in absolute value by 1. Using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \Phi, ||\cdot||_{Q,2}\right) = O(1),$$

so that Φ easily satisfies the entropy growth condition. Now let j index a generic moment function:

$$m_j(Y_{-k}, Z, U, \theta) = \left(\mathbb{1}\{U_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\} - 0.5\right) \mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\},$$

and let $\mathcal{M}_j$ be the associated class of functions:

$$\mathcal{M}_j = \left\{m_j(\cdot, u, \theta) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} : (u, \theta) \in \mathcal{U} \times \Theta\right\}.$$

Note that the values (z', y'_{-k}) are not arguments of the function, but instead are associated with the index j. Since π_k takes values in the interval $[-1, 1]$, the class $\mathcal{M}_j$ is uniformly bounded. We claim that there exists no set of size 2 shattered by $\mathcal{M}_j$, implying $\mathcal{M}_j$ is a VC-subgraph class. We will prove this by way of contradiction. In particular, suppose that there exists two points (y_1, z_1) and (y_2, z_2), and values $t_1, t_2 \in \mathbb{R}$ such that:

$$\left| \left\{ \begin{bmatrix} \mathbb{1}\{m_j(y_1, z_1, u, \theta) \geq t_1\} \\ \mathbb{1}\{m_j(y_2, z_2, u, \theta) \geq t_2\} \end{bmatrix} : (u, \theta) \in \mathcal{U} \times \Theta \right\} \right| = 4. \tag{3.182}$$

In other words, we suppose the set $\{(y_1, z_1), (y_2, z_2)\}$ is shattered by $\mathcal{M}_j$, and that $t_1, t_2 \in \mathbb{R}$ witness the shattering. We have:

$$m_j(y_1, z_1, u, \theta) = \left(\mathbb{1}\{u_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\} - 0.5\right) \mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\},$$
$$m_j(y_2, z_2, u, \theta) = \left(\mathbb{1}\{u_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{L_0 \pi_k(z', y'_{-k}; \theta), 0\} - 0.5\right) \mathbb{1}\{z_{2,k} = z, y_{2,-k} = y_{-k}\}.$$

Now consider two cases:

1. $\mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\} = \mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\}$: In this case the two functions $m_j(y_1, z_1, u, \theta)$ and $m_j(y_2, z_2, u, \theta)$ are identical for all $(u, \theta) \in \mathcal{U} \times \Theta$. This means (3.182) is impossible, since at least one of the vectors $(1, 0)$ and $(0, 1)$ cannot be picked out by $\mathcal{M}_j$.

2. $\mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\} \neq \mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\}$: In this case at least one of the functions $m_j(y_1, z_1, u, \theta)$ or $m_j(y_2, z_2, u, \theta)$ is the zero function. Again, this means (3.182) is impossible. For example, if $m_j(y_1, z_1, u, \theta)$ is the zero function, then it is impossible for $\mathcal{M}_j$ to pick out both $(0, 0)$ and $(1, 0)$ or both $(0, 1)$ and $(1, 1)$.

Since (y_1, z_1) and (y_2, z_2) were arbitrary, we conclude that there exists no set of size 2 shattered by $\mathcal{M}_j$. This implies that $\mathcal{M}_j$ is a VC-subgraph class, and using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \mathcal{M}_j, ||\cdot||_{Q,2}\right) = O(1).$$

Thus, $\mathcal{M}_j$ easily satisfies the entropy growth condition. Finally, let j' index a generic moment function:

$$m_j(Y_{-k}, Z, U, \theta) = \left(0.5 - \mathbb{1}\{U_k \leq \pi_k(z', y'_{-k}; \theta)\} - \max\{-L_0 \pi_k(z', y'_{-k}; \theta), 0\}\right) \mathbb{1}\{Z_k = z, Y_{-k} = y_{-k}\},$$

and let $\mathcal{M}_{j'}$ be the associated class of functions:

$$\mathcal{M}_{j'} = \left\{m_{j'}(\cdot, u, \theta) : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R} : (u, \theta) \in \mathcal{U} \times \Theta\right\}.$$

A nearly identical argument as for $\mathcal{M}_j$ reveals that $\mathcal{M}_{j'}$ is a VC-subgraph class and thus trivially satisfies the entropy growth condition. We conclude using Theorem 3.4.1(ii) that our class of policies Γ is PAMPAC learnable with a rate of convergence of $O(n^{-1/2})$.

3.C.2 Example 2: Program Evaluation

Verification of Assumptions 3.2.1, 3.2.2 and 3.2.3

We will now proceed to verify Assumption 3.2.1, 3.2.2 and 3.2.3. First note that Assumption 3.2.1 is trivially satisfied, since the probability space $(\Omega, \mathfrak{A}, P)$ is complete, $\mathcal{U}$ is a compact subset of euclidean space, and

Θ is a Polish space; in particular, since $\mathcal{Z}$ (and thus also $\mathcal{X}$) is finite, $\mathcal{G}$ can be considered as the set of all positive measurable functions $g : \mathcal{Z} \to [0, 1]$, in which case each $g \in \mathcal{G}$ has an equivalent representation as a vector in $[0, 1]^{|\mathcal{Z}|}$. The same logic applies to each $t \in \mathcal{T}$. Next, let us recall the multifunction:

$$\boldsymbol{G}^{-}(Y, D, Z, \theta) := \text{cl}\left\{(U_0, U_1, U) \in \mathcal{U} : \begin{array}{l} Y = U_0(1 - D) + U_1 D, \\ D = \mathbb{1}\{g(Z) \geq U\} \end{array}\right\}. \tag{3.183}$$

Close inspection of this multifunction shows that:

$$\boldsymbol{G}^{-}(y, d, z, \theta) = \begin{cases} \{y\} \times [\underline{Y}, \overline{Y}] \times [g(z), 1], & \text{if } d = 0, \\ [\underline{Y}, \overline{Y}] \times \{y\} \times [0, g(z)], & \text{if } d = 1. \end{cases} \tag{3.184}$$

Now for any $(u_0, u_1, u) \in \mathcal{U}$ we have:

$$\begin{aligned} &d((u_0, u_1, u), \boldsymbol{G}^{-}(Y, D, Z, \theta)) \\ &= D \max\{|u_0 - Y|, g(Z) - u\} + (1 - D) \max\{|u_1 - Y|, Z - g(z)\}. \end{aligned} \tag{3.185}$$

Since $g \in \mathcal{G}$ is measurable by definition, from here it is easily verified that the distance above is measurable with respect to $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{D}) \otimes \mathfrak{B}(\mathcal{Z})$. Since $(u_0, u_1, u) \in \mathcal{U}$ was arbitrary, by the result of Himmelberg (1975) (see also Theorem 1.3.3 in Molchanov (2017)) this implies that $\boldsymbol{G}^{-}$ is an Effros-measurable multifunction, as desired. Modulo changes in notation, it is easily seen that the conditional distribution of the vector (U_0, U_1, U) given (Y, D, Z) satisfies (3.4) in Assumption 3.2.2 using the multifunction in (3.15) with $g(\cdot) = g_0(\cdot)$. Finally, note that all of the moment functions in the moment conditions (3.17) - (3.22) are measurable and bounded by 1, and the moment functions from the moment conditions in (3.23) and (3.24) are measurable and bounded by $\max\{|\underline{Y}|, |\overline{Y}|\}$.

Turning to the counterfactual domain, recall the multifunction:

$$\boldsymbol{G}^{\star}(Z, U_0, U_1, U, \theta, \gamma) := \left\{(Y^{\star}_{\gamma}, D^{\star}_{\gamma}) \in \mathcal{Y} \times \{0, 1\} : \begin{array}{l} Y^{\star}_{\gamma} = U_0(1 - D^{\star}_{\gamma}) + U_1 D^{\star}_{\gamma}, \\ D^{\star}_{\gamma} = \mathbb{1}\{g(\gamma(Z)) \geq U\} \end{array}\right\}. \tag{3.186}$$

Note here we take $\mathcal{Y}^{\star} = \mathcal{Y}$, although this is not necessary. Furthermore, close inspection of this multifunction shows that:

$$\boldsymbol{G}^{\star}(z, u_0, u_1, u, \theta, \gamma) = \begin{cases} (u_1, 1), & \text{if } u \leq g(\gamma(z)), \\ (u_0, 0), & \text{if } g(\gamma(z)) < u. \end{cases} \tag{3.187}$$

In this case, the counterfactual map in (3.26) is single-valued. In this case, Effros measurability is equivalent to the usual notion of measurability for functions, and measurability of $\boldsymbol{G}^{\star}$ follows from familiar arguments after noting that both g and γ are measurable functions. Finally, modulo changes in notation, it is easily seen that the conditional distribution of the vector $(Y^{\star}_{\gamma}, D^{\star}_{\gamma})$ given (Y, D, Z, U_0, U_1, U) satisfies (3.6) in Assumption 3.2.3 using the multifunction in (3.26) with $g(\cdot) = g_0(\cdot)$.

Verification of Assumption 3.3.1

First we focus on (3.17) - (3.20). Since these moments do not depend on $t \in \mathcal{T}$, to verify Assumption 3.3.1 it suffices to focus on the parameter $g \in \mathcal{G}$. From the moment conditions (3.17) and (3.18) we have:

$$g(z_0, x) = P(D = 1 | Z = z_0, X = x) \iff \begin{cases} \mathbb{E}[(D - g_0(z_0, x))\, \mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0 \\ \mathbb{E}[(g_0(z_0, x) - D)\, \mathbb{1}\{Z_0 = z_0, X = x\}] \leq 0 \end{cases}, \tag{3.188}$$

and from (3.19) and (3.20) we have:

$$g_0(z_0, x) = P(U \leq g_0(z_0, x) | X = x) \iff \begin{cases} \mathbb{E}[(\mathbb{1}\{U \leq g_0(z_0, x)\} - g_0(z_0, x))\, \mathbb{1}\{X = x\}] \leq 0 \\ \mathbb{E}[(g_0(z_0, x) - \mathbb{1}\{U \leq g_0(z_0, x)\})\, \mathbb{1}\{X = x\}] \leq 0 \end{cases}, \tag{3.189}$$

For notational simplicity, let $g_0(z) := g_0(z_0, x)$ for $z = (z_0, x)$. From (3.188) we see that $g_0(z)$ is point-identified. Define:

$$\mathcal{G}^* = \{g : (g, t) \in \Theta^* \text{ for some } t \in \mathcal{T}\}.$$

Then point-identification of g_0 implies that $\mathcal{G}^*$ is a singleton, and that for any $g \in \mathcal{G}$:

$$d(g, \mathcal{G}^*) = \max_{z \in \mathcal{Z}} |g(z) - g_0(z)|.$$

From here it is straightforward to use conditions (3.188) and (3.189) to argue that part (i) of Assumption 3.3.1 is satisfied with $C_1 = 1$ for any $\delta > 0$. In particular, suppose $g \notin \mathcal{G}^*$, and that $z^* \in \mathcal{Z}$ satisfies:

$$d(g, \mathcal{G}^*) = \max_{z \in \mathcal{Z}} |g(z) - g_0(z)| = |g(z^*) - g_0(z^*)|.$$

Without loss of generality, suppose that $g(z^*) > g_0(z^*)$. Then from (3.188) we have:

$$\mathbb{E}[(g_0(z^*) - D)\, \mathbb{1}\{Z = z^*\}] = 0 < \mathbb{E}[(g(z^*) - D)\, \mathbb{1}\{Z = z^*\}].$$

Thus:

$$\begin{aligned} \mathbb{E}[(g(z^*) - D)\, \mathbb{1}\{Z = z^*\}] &= \mathbb{E}[(g(z^*) - D)\, \mathbb{1}\{Z = z^*\}] - \mathbb{E}[(g_0(z^*) - D)\, \mathbb{1}\{Z = z^*\}] \\ &= g(z^*) - g_0(z^*) \\ &= |g(z^*) - g_0(z^*)| \\ &= d(g, \mathcal{G}^*). \end{aligned}$$

Now to complete the verification of part (i) of Assumption 3.3.1 we turn to (3.21) - (3.24), which can be written as:

$$\mathbb{E}\left[t(z_0, x) - \mathbb{1}\{Z = z_0, X = x\}\right] = 0, \qquad \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X}, \tag{3.190}$$

and:

$$\mathbb{E}\left[U_d \left(\mathbb{1}\{Z = z_0, X = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0, x) - \mathbb{1}\{X = x\} t(z_0, x)\right)\right] \leq 0,\ \forall z_0 \in \mathcal{Z}_0,\ x \in \mathcal{X},\ d \in \{0, 1\}. \tag{3.191}$$

Since these moments do not depend on $g \in \mathcal{G}$, to verify Assumption 3.3.1 for these moments it suffices to focus on the parameter $t \in \mathcal{T}$. Now define:

$$\mathcal{T}^* = \{t : (g,t) \in \Theta^* \text{ for some } g \in \mathcal{G}\}.$$

From (3.190) it is clear that t_0 is also point identified. Since g_0 is also point identified we have $\Theta^* = \{g_0\} \times \{t_0\}$. Because of this, we claim that it suffices to focus on the conditions from (3.190); indeed, $t \notin \mathcal{T}^* \iff t \neq t_0$ implies that $t \notin \mathcal{T}^*$ if and only if (3.190) is violated. Now consider any $t \notin \mathcal{T}^*$ and let (z_0^*, x^*) satisfy:

$$(z_0^*, x^*) = \arg\max_{z_0, x} |t(z_0, x) - t_0(z_0, x)|.$$

Without loss of generality we can suppose $t(z_0, x) > t_0(z_0, x)$. Then:

$$\begin{aligned}
\mathbb{E}\left[t(z_0^*, x^*) - \mathbb{1}\{Z = z_0^*, X = x^*\}\right] &= \mathbb{E}\left[t(z_0^*, x^*) - \mathbb{1}\{Z = z_0^*, X = x^*\}\right] - \mathbb{E}\left[t_0(z_0^*, x^*) - \mathbb{1}\{Z = z_0^*, X = x^*\}\right] \\
&= t(z_0^*, x^*) - t_0(z_0^*, x^*) \\
&= |t(z_0^*, x^*) - t_0(z_0^*, x^*)| \\
&= d(t, \mathcal{T}^*).
\end{aligned}$$

Combining everything, if $\mathcal{J}$ indexes all the moment constraints and if $\theta \notin \Theta^*$ with $\theta = (g,t)$, then we know:

$$\inf_{P_{U_0,U_1,U|Y,D,Z} \in \mathcal{P}_{U_0,U_1,U|Y,D,Z}(\theta)} \max_{j \in \mathcal{J}} |\mathbb{E}[m_j(y,d,z,u_0,u_1,u,\theta)]|_+ \geq \max\{d(g,\mathcal{G}^*), d(t,\mathcal{T}^*)\} \geq d(\theta, \Theta^*).$$

Conclude that Assumption 3.3.1 is satisfied with $C_1 = 1$ for any $\delta > 0$.

For part (ii) of Assumption 3.3.1, we claim that we can set $C_2 = 1$. To show why, we will apply Lemma 3.3.1 to our environment. First note that φ is the identity function when we are interested in $\mathbb{E}_P[Y_\gamma^\star]$. Thus $L_\varphi = 1$ in Lemma 3.3.1. Next, note from the definition of our support restrictions $\boldsymbol{G}^-$ and $\boldsymbol{G}^\star$ we can deduce that:

$$d((u_0,u_1,u), \boldsymbol{G}^-(y,d,z,\theta)) = \begin{cases} \max\{|u_0 - y|, |g(z) - u|_+\}, & \text{if } d = 0, \\ \max\{|u_1 - y|, |u - g(z)|_+\}, & \text{if } d = 1. \end{cases} \tag{3.192}$$

$$d((y^\star, d^\star), \boldsymbol{G}^\star(y,d,z,u_0,u_1,u,\theta,\gamma)) = \begin{cases} \max\{|u_0 - y|, |g(z) - u|_+\}, & \text{if } u > g(\gamma(z)), \\ \max\{|u_1 - y|, |u - g(z)|_+\}, & \text{if } u \leq g(\gamma(z)). \end{cases} \tag{3.193}$$

We now define the sets Θ^- and $\Theta^\star$ given in Lemma 3.3.1 in the context of this example:

$$\begin{aligned}
\Theta^-(y,d,z,u_0,u_1,u) \cap \Theta_\delta^* &:= \left\{\theta \in \Theta_\delta^* : (u_0,u_1,u) \in \boldsymbol{G}^-(y,d,z,\theta)\right\} \\
&= \begin{cases} \{\theta \in \Theta_\delta^* : g(z) \in [0,u]\}, & \text{if } d = 0 \text{ and } u_0 = y, \\ \{\theta \in \Theta_\delta^* : g(z) \in [u,1]\}, & \text{if } d = 1 \text{ and } u_1 = y, \\ \emptyset, & \text{otherwise,} \end{cases}
\end{aligned} \tag{3.194}$$

$$\Theta^\star(v,\gamma) \cap \Theta^*_\delta := \{\Theta^*_\delta \in \Theta : (y^\star, d^\star) \in \boldsymbol{G}^\star(y, d, z, u_0, u_1, u, \theta, \gamma)\}$$

$$= \begin{cases} \{\theta \in \Theta^*_\delta : g(\gamma(z)) \in [0, u]\}, & \text{if } d^* = 0 \text{ and } y^\star = u_0, \\ \{\theta \in \Theta^*_\delta : g(\gamma(z)) \in [u, 1]\}, & \text{if } d^* = 1 \text{ and } y^\star = u_1, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{3.195}$$

With these definitions, we have for any $\theta \in \Theta^*_\delta$:

$$d(\theta, \Theta^-(y, d, z, u_0, u_1, u) \cap \Theta^*_\delta) = \begin{cases} |g(z) - u|_+, & \text{if } d = 0 \text{ and } u_0 = y, \\ |u - g(z)|_+, & \text{if } d = 1 \text{ and } u_1 = y, \\ +\infty, & \text{otherwise,} \end{cases} \tag{3.196}$$

$$d(\theta, \Theta^\star(v,\gamma) \cap \Theta^*_\delta) = \begin{cases} |g(\gamma(z)) - u|_+, & \text{if } d^* = 0 \text{ and } y^\star = u_0, \\ |u - g(\gamma(z))|_+, & \text{if } d^* = 1 \text{ and } y^\star = u_1, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{3.197}$$

Combining (3.192) with (3.196) we can verify condition (3.50) with $\ell_1 = 1$. Furthermore, by combining (3.193) with (3.197) we can verify condition (3.51) with $\ell_2 = 1$. Applying Lemma 3.3.1 then yields the choice $C_2 = L_\varphi \max\{\ell_1, \ell_2\} = 1$, as claimed above. Note also that this value of C_2 works for any $\delta > 0$.

It thus suffices to set $\mu^* = 1$ in Theorem 3.3.1. Also, recall the moment functions for this example from equations (3.17) - (3.20). The Theorem then states that the lower and upper bounds on the closed convex hull of the identified set for $\mathbb{E}[Y^\star_\gamma]$ can be computed as the solutions to the problems (3.52) and (3.53). Intuitively, under the assumptions of the Theorem the infimum over $\theta \in \Theta$ and supremum over $\theta \in \Theta$ in problems (3.52) and (3.53) will be obtained at the value $\theta_0 \in \Theta$.

Verification of Learnability

We claim that Φ is a VC class with VC index of at most $|\mathcal{Z}| + 1$. To prove this, we must show that there exists no set of points $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ with $n = |\mathcal{Z}| + 1$ shattered by Φ. Let $t_1, \ldots, t_n$ be arbitrary real numbers. Now define the set:

$$B := \left\{ \begin{bmatrix} \mathbb{1}\{\ \mathbb{1}\{g(\gamma(z_1)) \geq u\}(u_1 - u_0) + u_0 \geq t_1\ \} \\ \mathbb{1}\{\ \mathbb{1}\{g(\gamma(z_2)) \geq u\}(u_1 - u_0) + u_0 \geq t_2\ \} \\ \vdots \\ \mathbb{1}\{\ \mathbb{1}\{g(\gamma(z_n)) \geq u\}(u_1 - u_0) + u_0 \geq t_n\ \} \end{bmatrix} : (u_0, u_1, u, \theta) \in \mathcal{U} \times \Theta \right\}.$$

If B contains the vector $b \in \{0,1\}^n$, then we say that Φ "picks out" b. It suffices to show that there always exists at least one vector $b \in \{0,1\}^n$ that Φ fails to pick out. Since $n > |\mathcal{Z}|$, there exists at least one $z \in \mathcal{Z}$ that appears twice in the set $\mathcal{Z}_n$. Thus there is some $i, j \in \{1, \ldots, n\}$ such that $z_i = z_j$. Then regardless of the values of (u_0, u_1, u, θ) we will always have:

$$\mathbb{1}\{g(\gamma(z_i)) \geq u\}(u_1 - u_0) + u_0 = \mathbb{1}\{g(\gamma(z_j)) \geq u\}(u_1 - u_0) + u_0.$$

We then have:

1. If $t_i = t_j$ then Φ fails to pick out any vector $b \in \{0,1\}^n$ with $b_i = 0$ and $b_j = 1$.

2. If $t_i < t_j$ then Φ fails to pick out any vector $b \in \{0,1\}^n$ with $b_i = 0$ and $b_j = 1$.

3. If $t_j < t_i$ then Φ fails to pick out any vector $b \in \{0,1\}^n$ with $b_i = 1$ and $b_j = 0$.

Since this covers all possibilities for $t_i, t_j \in \mathbb{R}$, we conclude that there always exists at least one binary vector that Φ fails to pick out, and thus Φ shatters no set of size $n = |\mathcal{Z}| + 1$. Now using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \Phi, ||\cdot||_{Q,2}\right) = O(1),$$

so that Φ easily satisfies the entropy growth condition. Now let j index a generic moment function:

$$m_j(D, Z, \theta) = (D - g(z_0, x))\, \mathbb{1}\{Z_0 = z_0, X = x\},$$

and let $\mathcal{M}_j$ be the associated class of functions:

$$\mathcal{M}_j = \left\{m_j(\cdot, \theta) : \{0,1\} \times \mathcal{Z} \to \mathbb{R} : \theta \in \Theta\right\}.$$

Note this class indexes the moment functions from the moment conditions (3.17). Also note that (z_0, x) are not arguments of the moment function, but are instead are associated with the index j.

We claim that there exists no set of size 3 shattered by $\mathcal{M}_j$, implying $\mathcal{M}_j$ is a VC-subgraph class. We will prove this by way of contradiction. In particular, suppose that there exists three points (d_1, z_1), (d_2, z_2), and (d_3, z_3) and values $t_1, t_2, t_3 \in \mathbb{R}$ such that:

$$\left| \left\{ \begin{bmatrix} \mathbb{1}\{m_j(d_1, z_1, \theta) \geq t_1\} \\ \mathbb{1}\{m_j(d_2, z_2, \theta) \geq t_2\} \\ \mathbb{1}\{m_j(d_3, z_3, \theta) \geq t_3\} \end{bmatrix} : \theta \in \Theta \right\} \right| = 8. \tag{3.198}$$

In other words, we suppose the set $\{(d_1, z_1), (d_2, z_2), (d_3, z_3)\}$ is shattered by $\mathcal{M}_j$, and that $t_1, t_2, t_3 \in \mathbb{R}$ witness the shattering. We have:

$$\begin{aligned} m_j(d_1, z_1, \theta) &= (d_1 - g(z_0, x))\, \mathbb{1}\{z_{1,0} = z_0, x_1 = x\}, \\ m_j(d_2, z_2, \theta) &= (d_2 - g(z_0, x))\, \mathbb{1}\{z_{2,0} = z_0, x_2 = x\} \\ m_j(d_3, z_3, \theta) &= (d_3 - g(z_0, x))\, \mathbb{1}\{z_{3,0} = z_0, x_3 = x\}. \end{aligned}$$

Now consider two cases:

1. $\mathbb{1}\{z_{1,0} = z_0, x_1 = x\} = \mathbb{1}\{z_{2,0} = z_0, x_2 = x\} = \mathbb{1}\{z_{3,0} = z_0, x_3 = x\} = 1$: Note that since $d_i \in \{0,1\}$, at least two functions $m_j(d_1, z_1, \theta)$, $m_j(d_2, z_2, \theta)$ and $m_j(d_3, z_3, \theta)$ are identical for all $\theta \in \Theta$. This means (3.198) is impossible. For instance, suppose that $m_j(d_1, z_1, \theta) = m_j(d_2, z_2, \theta)$. Then at least one of the vectors $(1,0,0)$ or $(0,1,0)$ cannot be picked out by $\mathcal{M}_j$.

2. Either $\mathbb{1}\{z_{1,k} = z, y_{1,-k} = y_{-k}\} = 0$ or $\mathbb{1}\{z_{2,k} = z, y_{2,-k} = y_{-k}\} = 0$ or $\mathbb{1}\{z_{3,k} = z, y_{3,-k} = y_{-k}\} = 0$: In this case at least one of the functions $m_j(d_1, z_1, \theta)$, $m_j(d_2, z_2, \theta)$ or $m_j(d_3, z_3, \theta)$ is equal to zero for all $\theta \in \Theta$. Again, this means (3.198) is impossible. For example, if $m_j(d_1, z_1, \theta)$ is the zero function, then it is impossible for $\mathcal{M}_j$ to pick out both $(0,0,0)$ and $(1,0,0)$.

Since (d_1, z_1), (d_2, z_2), and (d_3, z_3) were arbitrary, we conclude that there exists no set of size 3 shattered by $\mathcal{M}_j$. This implies that $\mathcal{M}_j$ is a VC-subgraph class, and using, for example, Theorem 2.6.7 in Van Der Vaart

and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \mathcal{M}_j, ||\cdot||_{Q,2}\right) = O(1).$$

Thus, $\mathcal{M}_j$ easily satisfies the entropy growth condition. Given the relation between the moment functions from (3.17) and (3.18), a nearly identical analysis holds for the moment functions from the moment conditions (3.18).

Now let j' index a generic moment function:

$$m_{j'}(X, U, \theta) = \left(\mathbb{1}\{U \leq g(z_0, x)\} - g(z_0, x)\right) \mathbb{1}\{X = x\},$$

and let $\mathcal{M}_{j'}$ be the associated class of functions:

$$\mathcal{M}_{j'} = \left\{m_{j'}(\cdot, u, \theta) : \mathcal{X} \to \mathbb{R} : (u, \theta) \in \mathcal{U} \times \Theta\right\}.$$

Note this class indexes the moment functions from the moment conditions (3.19). Also note that (z_0, x) are not arguments of the moment function, but are instead are associated with the index j'.

We claim that there exists no set of size 2 shattered by $\mathcal{M}_{j'}$, implying $\mathcal{M}_{j'}$ is a VC-subgraph class. We will prove this by way of contradiction. In particular, suppose that there exists two points x_1 and x_2, and values $t_1, t_2 \in \mathbb{R}$ such that:

$$\left|\left\{\begin{bmatrix}\mathbb{1}\{m_{j'}(x_1, u, \theta) \geq t_1\} \\ \mathbb{1}\{m_{j'}(x_2, u, \theta) \geq t_2\}\end{bmatrix} : (u, \theta) \in \mathcal{U} \times \Theta\right\}\right| = 4. \tag{3.199}$$

In other words, we suppose the set $\{x_1, x_2\}$ is shattered by $\mathcal{M}_{j'}$, and that $t_1, t_2 \in \mathbb{R}$ witness the shattering. We have:

$$\begin{aligned} m_{j'}(x_1, u, \theta) &= \left(\mathbb{1}\{u \leq g(z_0, x)\} - g(z_0, x)\right) \mathbb{1}\{x_1 = x\}, \\ m_{j'}(x_2, u, \theta) &= \left(\mathbb{1}\{u \leq g(z_0, x)\} - g(z_0, x)\right) \mathbb{1}\{x_2 = x\}. \end{aligned}$$

Now consider two cases:

1. $\mathbb{1}\{x_1 = x\} = \mathbb{1}\{x_2 = x\} = 1$: Then the two functions $m_{j'}(x_1, u, \theta)$ and $m_{j'}(x_2, u, \theta)$ are identical for all $(u, \theta) \in \mathcal{U} \times \Theta$. This means (3.199) is impossible, since at least one of the vectors $(1, 0)$ and $(0, 1)$ cannot be picked out by $\mathcal{M}_{j'}$.

2. Either $\mathbb{1}\{x_1 = x\} = 0$ or $\mathbb{1}\{x_2 = x\} = 0$: In this case at least one of the functions $m_{j'}(x_1, u, \theta)$ or $m_{j'}(x_2, u, \theta)$ is the zero function. Again, this means (3.198) is impossible. For example, if $m_{j'}(x_1, u, \theta)$ is the zero function, then it is impossible for $\mathcal{M}_{j'}$ to pick out both $(0, 0)$ and $(1, 0)$.

Since x_1 and x_2 were arbitrary, we conclude that there exists no set of size 2 shattered by $\mathcal{M}_{j'}$. This implies that $\mathcal{M}_{j'}$ is a VC-subgraph class, and using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \mathcal{M}_{j'}, ||\cdot||_{Q,2}\right) = O(1).$$

Thus, $\mathcal{M}_{j'}$ easily satisfies the entropy growth condition. Given the relation between the moment functions from (3.19) and (3.20), a nearly identical analysis holds for the moment functions from the moment conditions (3.20).

Now let j'' index a generic moment function:

$$m_{j''}(Z,\theta) = t(z_0,x) - \mathbb{1}\{Z_0 = z_0, X = x)\},$$

and let $\mathcal{M}_{j''}$ be the associated class of functions:

$$\mathcal{M}_{j''} = \{m_{j''}(\cdot,\theta) : \mathcal{Z} \to \mathbb{R} : \theta \in \Theta\}.$$

Note this class indexes the moment functions from the moment conditions (3.21). Also note that (z_0, x) are not arguments of the moment function, but are instead are associated with the index j''.

We claim that there exists no set of size 3 shattered by $\mathcal{M}_{j''}$, implying $\mathcal{M}_{j''}$ is a VC-subgraph class. To see this, note that for any three points $\{z_1, z_2, z_3\}$ we have:

$$\begin{aligned}
m_{j''}(z_1,\theta) &= t(z_0,x) - \mathbb{1}\{z_{1,0} = z_0, x_1 = x)\},\\
m_{j''}(z_2,\theta) &= t(z_0,x) - \mathbb{1}\{z_{2,0} = z_0, x_2 = x)\},\\
m_{j''}(z_3,\theta) &= t(z_0,x) - \mathbb{1}\{z_{2,0} = z_0, x_2 = x)\}.
\end{aligned}$$

The conclusion follows from the fact that two of these moment functions must always be the same. This implies that $\mathcal{M}_{j''}$ is a VC-subgraph class, and using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \mathcal{M}_{j''}, ||\cdot||_{Q,2}\right) = O(1).$$

Thus, $\mathcal{M}_{j''}$ easily satisfies the entropy growth condition. Given the relation between the moment functions from (3.21) and (3.22), a nearly identical analysis holds for the moment functions from the moment conditions (3.22).

Finally, let j''' index a generic moment function:

$$m_{j'''}(Z, U_d, \theta) = U_d \left(\mathbb{1}\{Z = z_0, X = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0,x) - \mathbb{1}\{X = x\} t(z_0,x) \right).$$

and let $\mathcal{M}_{j''}$ be the associated class of functions:

$$\mathcal{M}_{j'''} = \left\{ m_{j'''}(\cdot, u_d, \theta) : \mathcal{Z} \to \mathbb{R} : (u_d, \theta) \in [\underline{Y}, \overline{Y}] \times \Theta \right\}.$$

Note this class indexes the moment functions from the moment conditions (3.23). Also note that (z_0, x) are not arguments of the moment function, but are instead are associated with the index j'''.

We claim that there exists no set of size 5 shattered by $\mathcal{M}_{j'''}$, implying $\mathcal{M}_{j'''}$ is a VC-subgraph class. To see this, note that for any five points $\{z_1, z_2, z_3, z_4, z_5\}$ we have:

$$\begin{aligned}
m_{j'''}(z_1, u_d, \theta) &= u_d \left(\mathbb{1}\{z_{1,0} = z_0, x_1 = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0,x) - \mathbb{1}\{x_1 = x\} t(z_0,x) \right),\\
m_{j'''}(z_2, u_d, \theta) &= u_d \left(\mathbb{1}\{z_{2,0} = z_0, x_2 = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0,x) - \mathbb{1}\{x_2 = x\} t(z_0,x) \right),\\
m_{j'''}(z_3, u_d, \theta) &= u_d \left(\mathbb{1}\{z_{3,0} = z_0, x_3 = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0,x) - \mathbb{1}\{x_3 = x\} t(z_0,x) \right),
\end{aligned}$$

$$m_{j'''}(z_4, u_d, \theta) = u_d \left(\mathbb{1}\{z_{4,0} = z_0, x_4 = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0, x) - \mathbb{1}\{x_4 = x\} t(z_0, x) \right),$$
$$m_{j'''}(z_5, u_d, \theta) = u_d \left(\mathbb{1}\{z_{5,0} = z_0, x_5 = x\} \sum_{z_0 \in \mathcal{Z}_0} t(z_0, x) - \mathbb{1}\{x_5 = x\} t(z_0, x) \right).$$

The conclusion follows from the fact that two of these moment functions must always be identical for all θ. This implies that $\mathcal{M}_{j'''}$ is a VC-subgraph class, and using, for example, Theorem 2.6.7 in Van Der Vaart and Wellner (1996), we can deduce:

$$\sup_{Q \in \mathcal{Q}_n} \log N\left(\varepsilon, \mathcal{M}_{j'''}, ||\cdot||_{Q,2}\right) = O(1).$$

Thus, $\mathcal{M}_{j'''}$ easily satisfies the entropy growth condition. Given the relation between the moment functions from (3.23) and (3.24), a nearly identical analysis holds for the moment functions from the moment conditions (3.24).

Combining everything and applying Theorem 3.4.1(ii), we thus have that the policy space Γ for this problem is learnable.

Bibliography

Aliprantis, C. D. and Border, K. C. (2006). *A Hitchhiker's Guide to Infinite dimensional analysis*. Springer.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631.

Andrews, D. W. and Guggenberger, P. (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory*, 25(3):669–709.

Andrews, D. W. and Kwon, S. (2019). Inference in moment inequality models that is robust to spurious precision under model misspecification. Working paper.

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

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

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

Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 2(4):343–370.

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

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

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.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1-3):85–113.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Long, P. M., and Williamson, R. C. (1996). Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452.

Belloni, A., Bugni, F., and Chernozhukov, V. (2018). Subvector inference in partially identified models with many moment inequalities. *arXiv preprint arXiv:1806.11466*.

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

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

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.

Bertsekas, D. P. and Shreve, S. (1978). *Stochastic optimal control: the discrete-time case*. Academic Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.

Boozer, M. and Cacciola, S. E. (2001). Inside the 'black box' of project star: Estimation of peer effects using experimental data. *Yale Economic Growth Center Discussion Paper*, (832).

Borwein, J. and Lewis, A. S. (2010). *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

Bousquet, O., Koltchinskii, V., and Panchenko, D. (2002). Some local measures of complexity of convex hulls and generalization bounds. In *International Conference on Computational Learning Theory*, pages 59–73. Springer.

Boyd-Zaharias, J., Finn, J., Fish, R., and Gerber, S. (2007). Project star and beyond: Database users guide. *HEROS Inc. and University of New York at Buffalo technical report*.

Bresnahan, T. F. and Reiss, P. C. (1990). Entry in monopoly market. *The Review of Economic Studies*, 57(4):531–553.

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

Brock, W. A. and Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260.

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

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

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81.

Chamberlain, G. (2011). Bayesian aspects of treatment choice. *The Oxford Handbook of Bayesian Econometrics*, pages 11–39.

Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

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

Chernozhukov, V., Hong, H., and Tamer, E. (2007b). 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. (2015). Characterizations of identified sets delivered by structural econometric models. Technical report, cemmap working paper.

Chesher, A. and Rosen, A. M. (2012). Simultaneous equations models for discrete outcomes: coherence, completeness, and identification. Technical report, CEMMAP working paper.

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

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

Chesher, A. and Rosen, A. M. (2017b). Incomplete english auction models with heterogeneity. Technical report, cemmap working paper.

Chesher, A. and Rosen, A. M. (2020). Structural modeling of simultaneous discrete choice. Working paper.

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

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

Cho, J. and Russell, T. M. (2019). Simple inference on functionals of set-identified parameters defined by linear moments. *arXiv preprint arXiv:1810.03180*.

Ciliberto, F., Murry, C., and Tamer, E. T. (2018). Market structure and competition in airline markets. *Available at SSRN 2777820*.

Cohn, D. L. (2013). *Measure theory*. Springer.

Corbae, D., Stinchcombe, M. B., and Zeman, J. (2009). *An introduction to mathematical analysis for economic theory and econometrics*. Princeton University Press.

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

Dolgopolik, M. (2016). A unifying theory of exactness of linear penalty functions. *Optimization*, 65(6):1167–1202.

Dontchev, A. L. and Rockafellar, R. T. (2009). Implicit functions and solution mappings. *Springer Monogr. Math.*

Dudley, R. M. (2010). *Real analysis and probability*. Cambridge university press.

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

Dudley, R. M., Giné, E., and Zinn, J. (1991). Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510.

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

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

Firpo, S. and Ridder, G. (2019). Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234.

Freyberger, J. and Horowitz, J. L. (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, 189(1):41–53.

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

Gafarov, B., Meier, M., and Olea, J. L. M. (2018). Delta-method inference for a class of set-identified SVARs. *Journal of Econometrics*, 203(2):316–327.

Galichon, A. and Henry, M. (2006). Inference in incomplete models. *Available at SSRN 886907*. Working paper.

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*, 78(4):1264–1298.

Giné, E., Koltchinskii, V., et al. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216.

Giné, E., Koltchinskii, V., and Wellner, J. A. (2003). Ratio limit theorems for empirical processes. In *Stochastic inequalities and applications*, pages 249–278. Springer.

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

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, 111(1):1–51.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature*, 48(2):356–98.

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. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734.

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.

Himmelberg, C. (1975). Measurable relations. *Fundamenta Mathematicae*, 87(1):53–72.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

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

Hoffmann-Jørgensen, J. (1991). *Stochastic processes on Polish spaces*. Various publications series. Aarhus Universitet. Matematisk Institut.

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.

Hurwicz, L. (1950). Generalization of the concept of identification. *Statistical inference in dynamic economic models*, 10:245–57.

Ichimura, H. and Taber, C. R. (2000). Direct estimation of policy impacts. Working paper.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Ioffe, A. (2016). Metric regularity—a survey part 1. theory. *Journal of the Australian Mathematical Society*, 101(2):188–243.

Jia, P. (2008). What happens when wal-mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica*, 76(6):1263–1316.

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. (2019a). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Kaido, H., Molinari, F., and Stoye, J. (2019b). Constraint qualifications in partial identification. *arXiv preprint arXiv:1908.09103*.

Kaido, H. and Santos, A. (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica*, 82(1):387–413.

Kalouptsidi, M., Kitamura, Y., Lima, L., and Souza-Rodrigues, E. (2019). Partial identification and inference for dynamic models and counterfactuals. Working paper.

Kasy, M. (2016). Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 98(1):111–131.

Kasy, M. (2019). Uniformity and the delta method. *Journal of Econometric Methods*, 8(1).

Kearns, M. J. and Schapire, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497.

Kearns, M. J., Vazirani, U. V., and Vazirani, U. (1994). *An introduction to computational learning theory*. MIT press.

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

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer.

Koopmans, T. C., Rubin, H., and Leipnik, R. B. (1950). Measuring the equation systems of dynamic economics. *Statistical inference in dynamic economic models*, 10.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

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.

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

Lafférs, L. (2013b). Identification in models with discrete variables. Working paper.

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

Lewbel, A. (2007). Coherency and completeness of structural models containing a dummy endogenous variable. *International Economic Review*, 48(4):1379–1392.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, pages 455–476.

Li, L. (2019). Identification of structural and counterfactual parameters in a large class of structural econometric models. Working paper.

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.

Luo, Z.-Q., Pang, J.-S., and Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge University Press.

Manski, C. and Tetenov, A. (2014). The quantile performance of statistical treatment rules using hypothesis tests to allocate a population to two treatments. Technical report, cemmap working paper.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, 25(1):79–104.

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

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

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. (2011). Actualist rationality. *Theory and Decision*, 71(2):195–210.

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

Marshak, J. (1953). Economic measurements for policy and prediction. *Studies in Econometric Method*, pages 1–26.

Massart, P. (2000). Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303.

Mbakop, E. and Tabord-Meehan, M. (2019). Model selection for treatment choice: Penalized welfare maximization. *arXiv preprint arXiv:1609.03167*.

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

Miyauchi, Y. (2016). Structural estimation of pairwise stable networks with nonnegative externality. *Journal of econometrics*, 195(2):224–235.

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

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

Molchanov, I. (2017). *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.

Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge University Press.

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

Mourifie, I., Henry, M., and Méango, R. (2018). Sharp bounds and testability of a roy model of stem major choices. *Available at SSRN 2043117*.

Mourifie, I. and Wan, Y. (2020). Layered sensitivity analysis in program evaluation using the mte. Working paper.

Munkres, J. (2014). *Topology*. Pearson Education.

Ok, E. A. (2007). *Real analysis with economic applications*, volume 10. Princeton University Press.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2011). Moment inequalities and their applications, discussion paper, harvard.

Pang, J.-S. (1997). Error bounds in mathematical programming. *Mathematical Programming*, 79(1-3):299–332.

Parthasarathy, K. R. (2005). *Probability measures on metric spaces*, volume 352. American Mathematical Soc.

Pearl, J. (2009). *Causality: models, reasoning and inference*. Springer.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR.

Qin, D. and Gilbert, C. L. (2001). The error term in the history of time series econometrics. *Econometric theory*, 17(2):424–450.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138(9):2786–2807.

Rostek, M. (2010). Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1):339–371.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Russell, T. (2019). Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Available at SSRN 3013430*.

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

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66(3):477–487.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

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

Shreve, S. E. and Bertsekas, D. P. (1978). Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978.

Shreve, S. E. and Bertsekas, D. P. (1979). Universally measurable policies in dynamic programming. *Mathematics of Operations Research*, 4(1):15–30.

Spingarn, J. E. and Rockafellar, R. T. (1979). The generic nature of optimality conditions in nonlinear programming. *Mathematics of Operations Research*, 4(4):425–430.

Stinchcombe, M. B. and White, H. (1992). Some measurability results for extrema of random functions over random sets. *The Review of Economic Studies*, 59(3):495–514.

Stoye, J. (2009a). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

Stoye, J. (2009b). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

Stoye, J. (2011). Statistical decisions under ambiguity. *Theory and decision*, 70(2):129–148.

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156.

Syrgkanis, V., Tamer, E., and Ziani, J. (2018). Inference on auctions with weak assumptions on information. *arXiv preprint arXiv:1710.03830*.

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

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

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165.

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

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

Uetake, K. and Watanabe, Y. (2019). Entry by merger: Estimates from a two-sided matching model with externalities. *Available at SSRN 2188581*.

Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ).

Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. (1998). Statistical learning theory. *New York*.

Vidyasagar, M. (2002). *A theory of learning and generalization*. Springer-Verlag.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341.

Wachsmuth, G. (2013). On LICQ and the uniqueness of lagrange multipliers. *Operations Research Letters*, 41(1):78–80.

Wald, A. (1950). *Statistical decision functions*. Wiley.

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