Long-Term Treatment Effect Identification via Data Combination

Filip Obradović*

(Preliminary and Incomplete. Please do not circulate.)

Final Version Will be Posted Here

October 15, 2024

Abstract

Recent literature proposes combining short-term experimental and long-term observational data to provide credible alternatives to observational studies for identification of long-term average treatment effects (LTEs). The paper makes two distinct contributions in this context. I first show that experimental data have an *auxiliary role*. They bring *no identifying power* without additional assumptions in the observational data. When such assumptions are imposed, experimental data serve to *amplify* their power. If the assumptions fail, adding experimental data may only lead to results that are farther from the truth. Motivated by this, I introduce two assumptions on *treatment response* that may be defensible based on economic theory or intuition. To utilize them, I develop a novel two-step identification approach that centers on bounding *temporal link functions* – the relationship between short-term and mean long-term potential outcomes. This approach allows for imperfect experimental compliance and provides sharp bounds on LTEs for a general class of assumptions – extending existing results under compliance issues. I illustrate the findings by studying the long-term effects of Head Start participation. To do so, I create a new combined dataset using the Head Start Impact Study and the NLSY79 Child and Young Adult cohort.

*Northwestern University, Department of Economics. Email: obradovicfilip@u.northwestern.edu

I am deeply grateful to Charles Manski, Ivan Canay and Federico Bugni for their guidance and support. I am also thankful to Joel Horowitz, Eric Auerbach and Piotr Dworczak for valuable suggestions. I thank the participants at the Econometrics Seminar, Econometrics Reading Group at Northwestern, and the 2024 Junior Econometrics Conference at Notre Dame for comments. Financial support from the Robert Eisner Memorial Fellowship and the Unicredit Crivelli Scholarship is gratefully acknowledged.

1 Introduction

Identification of the long-term average treatment effect (henceforth LTE) is an important goal in economics and various other fields of science. For example, one may be interested in the effects of childhood intervention on outcomes in adulthood; impact of conditional cash transfers early in life on employment prospects; or adverse/protective effects of vaccination long after treatment. Gupta et al. (2019) explain that identifying the LTE is also recognized as an important challenge by researchers in the private sector.

Identification of the LTE is commonly done using observational data (for examples, see Currie and Almond (2011), Hoynes and Schanzenbach (2018)). However, observational studies critically rely on assumptions that may often be deemed implausible, typically those restricting the treatment selection mechanism. While randomized controlled trials (RCTs) eliminate the need for such assumptions, long-term experiments may be prohibitively costly or infeasible.[1] Short-term RCTs may be more feasible, but they do not reveal the long-term outcomes and hence the LTE. Nevertheless, short-term RCTs may complement observational data.

Motivated by this, a large body of recent work following Athey, Chetty, and Imbens (2020) and Athey et al. (2024) aims to provide credible alternatives to conventional observational studies that rely on a combination of: 1) a long-term observational dataset with non-randomized treatment assignment; 2) a short-term experimental dataset with unobserved long-term outcomes.[2] Pursuing point identification, this literature commonly imposes assumptions on the selection mechanism in the observational data, mirroring conventional observational studies. Ghassami et al. (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023) and Imbens et al. (2024) argue that frequently used assumptions may fail in contexts of economic interest; Park and Sasaki (2024a) show that they generally fail in common selection models, including the Roy model. It is more broadly acknowledged that selection assumptions can be challenging to justify. However, existing results do not reveal whether the addition of experimental data provides identifying power in absence of such assumptions; or perhaps lessens the importance of their plausibility.

This paper shows that neither is true. It demonstrates that experimental data bring no identifying power in the absence of assumptions in the observational data, nor do they diminish the importance of their plausibility; in fact, they may even exacerbate it, as will soon be made precise. Building on this, the paper introduces two assumptions on *treatment response* that may be defensible based on economic theory or intuition, instead of selection assumptions. To utilize

1. Institutions supporting RCTs in development economics frequently require phase-in designs with staggered rollout of treatment to the whole sample. This limits follow-up for the control group.

2. Structural modeling with data combination predates this work (Todd and Wolpin (2006), Attanasio, Meghir, and Santiago (2012), García et al. (2020), Todd and Wolpin (2023)). The focus here is on "reduced form" methods.

them, I develop a novel two-step identification approach that bounds *temporal link functions* – means of long-term potential outcomes conditional on short-term potential outcome – as an intermediate step. Like related work, I center on enabling plausible inference. I do so by providing defensible assumptions in the observational data and allowing for imperfect experimental compliance in the framework. I now expand upon these points.

Due to the possibility of imperfect compliance and weak assumptions imposed in observational data, the LTE will generally be partially identified. The set of all values of the LTE that are consistent with data and assumptions is called the identified set. The identification results yield the smallest such set, or the so-called sharp identified set. While the identification analysis generally produces bounds on the LTE, it does not preclude point identification which may occur under some data generating processes.

As the first contribution, I uncover the role of the experimental dataset and its implications. Experimental data provide no identifying power, per se; the identified sets for the LTE obtained from combined and solely observational data are equal in the absence of additional assumptions in the observational data.[3] Additional assumptions in the observational data are thus necessary to leverage the experimental data and to identify the sign of the LTE. When the assumptions are imposed, the identified set based on combined data is a subset of the one that uses only observational data. It need not be a strict subset. Hence, the experimental data serve to potentially, but not necessarily, amplify the identifying power of the assumptions. These observations reveal the *auxiliary* role of experimental data. Assumptions in the observational data remain *central* under data combination, mirroring their prominence in observational studies.

I illustrate this point via a widely used selection assumption in this literature - latent unconfoundedness (LUC). I show that LUC may commonly have identifying power for the LTE using observational data alone. When experimental data is added, LUC has more identifying power and point identifies the LTE. In extreme cases, observational data may even point identify the LTE under LUC without requiring experimental data. Since the LTE is point identified from combined data and LUC, the experimental data potentially, but not necessarily, amplify the identifying power of LUC.

Due to the amplifying role of the experimental data, data combination does not lessen the importance of plausible assumptions in the observational data. On the contrary, it may only exacerbate it. When these assumptions are implausible, it may be preferable to discard the experimental data. Under misspecified assumptions, the identified set obtained using combined data can never be closer to the true LTE than the set obtained using only observational data. This finding is an application of a more general lemma. The lemma states that whenever two

3. This result may resemble findings of Park and Sasaki (2024b). Remark 2 explains that our conclusions differ.

misspecified identified sets are nested, the smaller one must be at least as far from the truth as the larger one. This result is strikingly simple, but it appears that it was not formalized before.

For the second contribution, I introduce two assumptions on temporal link functions – *latent monotone instrumental variable (LIV)* and *treatment invariance (TI)*. LIV asserts that the temporal link functions are non-decreasing. In other words, by LIV means of long-term potential outcomes are non-decreasing in short-term potential outcomes. It is related to the *monotone instrumental variable* assumption of Manski and Pepper (2000). Here, the instrumental variable is the latent potential outcome. TI posits that the temporal link functions are invariant to the treatment, or that the relationship between short-term potential outcome and the mean long-term potential outcome is unaffected by the treatment.

While LIV may be argued based on intuition, TI is implied by a model. For example, TI would hold under the model proposed by García et al. (2020) in the context of early childhood intervention. LIV and TI do not impose any restrictions on the selection mechanism and represent assumptions on treatment response. In contrast, the existing literature primarily imposes restrictions on the selection mechanism. Manski (1997) notes that convincing behavioral arguments are often lacking for such assumptions.

The third contribution is a novel two-step identification approach that enables the use of proposed assumptions. To summarize the main idea, let $Y(d)$ and $S(d)$ denote long- and short-term potential outcomes under treatment $d \in \{0, 1\}$ and let:

$$\begin{aligned} m_d(s) &:= E[Y(d)|S(d) = s]; \\ \gamma_d &:= P_{S(d)}. \end{aligned}$$

I refer to $m_d(s)$ as the *temporal link functions*, while γ_d are the distributions of short-term potential outcomes. We can then write the LTE using the identity:

$$LTE := E[Y(1) - Y(0)] = \underbrace{\int_{\mathcal{S}} m_1(s) d\gamma_1(s)}_{E[Y(1)]} - \underbrace{\int_{\mathcal{S}} m_0(s) d\gamma_0(s)}_{E[Y(0)]}, \tag{1}$$

In the first step, I find all $(m_0, m_1, \gamma_0, \gamma_1)$ compatible with the data and assumptions, which then yield all feasible values for the LTE in the second step.

For the first step, I derive the sharp identified set for $(m_0, m_1, \gamma_0, \gamma_1)$ under a generic restriction on (m_0, m_1) which embodies the assumptions in the observational data. Sharp characterization requires joint identification of $(m_0, m_1, \gamma_0, \gamma_1)$ due to cross-restrictions imposed by the data. Intuitively, these stem from the fact that m_d conditions on a latent random variable $S(d)$, which, in turn, determines the distribution γ_d. To tractably summarize all information contained in the

data and assumptions, I combine two concepts from random set theory – Artstein's inequalities and the conditional Aumann expectation. In the second step, I collect all values that possible $(m_0, m_1, \gamma_0, \gamma_1)$ produce via (1), which yields the identified set for the LTE. I show that this set can be characterized as an interval bounded by two bilevel optimization problems in commonly considered settings by the literature. To alleviate the computational burden, I demonstrate that the inner optimization problems have closed-form solutions under LIV and TI. I further reduce the number of constraints in the outer optimization problems via the concept of *core determining classes* (Galichon and Henry (2011)).

The identification approach has several appealing features. First, it produces sharp bounds on the LTE for any assumption that can be represented as a restriction on (m_0, m_1), without requiring individual proofs of sharpness. This is a commonly exploited benefit of utilizing tools from random set theory. For similar results in other models see Chesher and Rosen (2017), Russell (2021), Park and Sasaki (2024b), and Han and Kaido (2024). Consequently, this: 1) yields the desired results for proposed assumptions LIV and TI; 2) allows researchers to easily characterize identified sets under new assumptions tailored to their empirical setting. Second, the approach can accommodate imperfect compliance in the experimental data by allowing partial identification of the subvector (γ_0, γ_1). Third, it directly extends previously proposed results. The approach allows existing identification strategies to account for imperfect compliance since they can be represented as restrictions on (m_0, m_1), .

Accommodating imperfect compliance is of great practical relevance. Compliance issues are prevalent in RCTs, especially in the experiments previously used in this context.[4] Moreover, parameters such as the intent-to-treat effect (ITT) and local average treatment effects (LATE) of Imbens and Angrist (1994) are unidentified in this setting unlike in classical RCTs. This is because the long-term outcomes and randomized treatment assignment are never observed simultaneously. Even if possible, this entails identifying a parameter other than the LTE, which may or may not be of interest depending on the research question. For more details see discussions in Deaton (2009), Heckman and Urzua (2010) and Imbens (2010). In spite of this, to the extent of my knowledge, related literature did not consider experiments with imperfect compliance.

Incomplete Results: To illustrate the utility of the results, I create a new combined dataset using individuals in the Head Start Impact Study (HSIS) and the National Longitudinal Survey of Youth NLSY79 Child and Young Adult (CNLSY). I estimate the treatment effects of Head Start participation on a variety of outcomes in adulthood, utilizing childhood cognitive test scores as short-term outcomes. Quantification of long-term treatment effects of Head Start has

4. Athey, Chetty, and Imbens (2020) and Park and Sasaki (2024a) use the Project STAR and Aizer et al. (2024) the Job Corps RCT. Both had significant reassignment/compliance issues. (e.g. see Chen, Flores, and Flores-Lagunes (2018) and Russell (2021)) Non-compliance rate is 16% in the empirical illustration.

long been an active field of research with important policy applications. Yet, due to lack of consensus on the plausible approach to identification, agreement on the sign and magnitude of the effects has not been reached (see the review in Pages et al. (2020)). While the results in this paper mainly represent an illustrative application, they show that data combination may provide fruitful contributions to the discourse on this long-standing open question.

Section 2 introduces the setting and identifying assumptions. Section 3 studies the role of the experimental data. Section 4 develops the identification framework. Section 5 provides a tractable characterization of the identified set and a consistent estimator. Section 6 applies the method to study Head Start participation. Section 7 concludes. Appendix A contains the extensions of the findings, and Appendix B lists previously known supporting results. Appendix C collects the proofs.

2 Setting

I formalize the problem using the standard potential outcomes model. Let $Y(d) \in \mathcal{Y}$ and $S(d) \in \mathcal{S}$ denote the long-term and short-term potential outcomes under some binary treatment $d \in \{0,1\}$, respectively.[5] Denote the realized treatment by $D \in \{0,1\}$. The observed outcomes are:

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ S &= DS(1) + (1-D)S(0). \end{aligned} \tag{2}$$

The notation implies that there are no spillovers, or the standard stable unit treatment value assumption (SUTVA). Let $X \in \mathcal{X}$ be a vector of observed covariates. Define the conditional *long-term* average treatment effect (CLTE) $\tau(x)$:

$$\tau(x) = E[Y(1) - Y(0)|X = x]. \tag{3}$$

The parameter of interest can be the CLTE itself or its weighted averages - average *long-term* treatment effect (LTE). I focus on the former for generality and discuss the latter in Section A.2. Throughout the paper, I assume $E[|Y(d)|] < \infty$ for $d \in \{0,1\}$, which ensures that the parameters are well defined.

Example 1. *(Head Start Participation)* In the illustrative application, D is an indicator for Head Start participation, $S(d)$ are cognitive test scores in childhood, and $Y(d)$ are outcomes in adulthood, such as earnings, under treatment d.

5. Supports are invariant to the treatment. This can be relaxed at the expense of more complicated notation.

2.1 Observed Data

As in Athey, Chetty, and Imbens (2020), I maintain the existence of a population divided into two subpopulations from which the two datasets are randomly drawn: a short-term experimental and a long-term observational dataset. Let $G \in \{O, E\}$ be the indicator for the subpopulation, where $G = O$ generates the observational and $G = E$ the experimental dataset.[6]

Let $Z \in \mathcal{Z}$ be an exogenous (i.e. randomly assigned) instrument in the experiment, inducing individuals into treatment. In the experimental dataset, the researcher observes (S, D, X, Z), but not Y. For individuals in the observational dataset, (Y, S, D, X) are observed. Note the absence of Z when $G = O$, as I do not assume the existence of an exogenous instrument in the observational data.

Z serves to allow for imperfect experimental compliance. Usually, $Z \in \{0, 1\}$, representing random assignment to the treatment or the control group. The identification analysis can accommodate bounded $\mathcal{Z}$ with multiple or even a continuum of points $\mathcal{Z} = [0, 1]$, as in Heckman and Vytlacil (1999). For expositional simplicity, I refer to experiments with $P(D = Z|G = E) < 1$ as having imperfect compliance, as opposed to perfect compliance when $P(D = Z|G = E) = 1$. I thus also refer to Z as treatment assignment regardless of its support, keeping in mind that the $\mathcal{Z}$ may contain points beyond $\{0, 1\}$.

Example 1 (continued). The observational dataset is the National Longitudinal Survey of Youth (NLSY), and the experimental dataset is the Head Start Impact Study (HSIS). In the HSIS, $Z = 1$ if the individual is assigned to participation in Head Start and $Z = 0$ if assigned to non-participation. D is the indicator for true participation. Kline and Walters (2016) explain that some individuals may have $D \neq Z$.

A distinguishing feature of this paper is that imperfect experimental compliance is possible. This is important for two reasons. First, noncompliance is prevalent in practice. Second, the intent-to-treat effects (ITT) or the local average treatment effect (LATE) are unidentified in this setting. Researchers often obviate compliance issues by focusing on ITT and LATE that remain identified in conventional experiments. Identification of ITT and LATE requires jointly observing treatment *assignment* Z and the long-term outcomes Y. Since Z is never jointly observed with Y in this setting, both parameters are unidentified. touch up

I maintain the following assumptions throughout the paper.

Assumption RA. *(Random Assignment)* $Z \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = E$

Assumption EV. *(Experimental External Validity)* $G \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X$.

6. This setting has become common. See also García et al. (2020), Athey, Chetty, and Imbens (2020), Ghassami et al. (2022), Hu, Zhou, and Wu (2022), Van Goffrier, Maystre, and Gilligan-Lee (2023), Chen and Ritzwoller (2023), Park and Sasaki (2024a), Aizer et al. (2024) and Imbens et al. (2024).

Assumption RA holds if Z in the experimental data is randomly assigned. It is a standard assumption in the program evaluation literature. $D \not\perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = g$ is permitted for any $g \in \{O, E\}$. Generally, this is expected in the observational dataset, i.e. for $g = O$. In the experimental data, it is also expected when there is imperfect compliance. When compliance is perfect, Assumption RA implies $D \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))|X, G = E$. I do not assume that $P(D = 1|G = g) \in (0, 1)$ for any $g \in \{0, 1\}$. Instead, $P(D = 1|G = g) \in [0, 1]$ which may be relevant for $g = O$ when a certain treatment is only available in the experiment. This is the case with some early childhood intervention programs or novel vaccines.

Assumption EV is a standard assumption in the data combination literature, linking the two datasets. It states that the subpopulations generating them do not differ in terms of counterfactual distributions (conditional on X). It holds when participants are randomly recruited into the datasets from the same population (conditional on X). Section A.3, I present the strongest testable implication.

Under Assumption EV, CLTE is invariant to G, $E[Y(1) - Y(0)|X = x, G] = E[Y(1) - Y(0)|X = x] = \tau(x)$. Henceforth, I keep conditioning on X implicit. The following analysis should be understood as conditional-on-X, and I write the parameter of interest $\tau(x)$ as:

$$\tau = E[Y(1) - Y(0)] \tag{4}$$

with the understanding that it refers to CLTE rather than LTE.

Notation: I denote laws $P(\cdot|\mathcal{E}, G = g)$ where $\mathcal{E}$ is some event as $P_g(\cdot|\mathcal{E})$ for $g \in \{O, E\}$. Whenever $P_E(\cdot|\mathcal{E}) = P_O(\cdot|\mathcal{E})$, I omit the subscript. This is inherited by their features $E[\cdot|\mathcal{E}, G = g] = E_g[\cdot|\mathcal{E}]$ and $V[\cdot|\mathcal{E}, G = g] = V_g[\cdot|\mathcal{E}]$. I also denote laws of random elements using subscripts when g is irrelevant, and the element needs to be specified (e.g. $P_{S(d)}$ is the law of $S(d)$).

2.2 Identification Preliminaries

This paper proposes a novel identification approach. To introduce it, recall that for $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$m_d(s) := E[Y(d)|S(d) = s] \tag{5}$$

$$\gamma_d := P_{S(d)}. \tag{6}$$

I refer to $m_d(s)$ as *temporal link functions*, since they "link" the short-term and long-term potential outcomes in a way that is meaningful for identification of τ. We can write the parameter

of interest as:

$$\tau = E[Y(1) - Y(0)] = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{7}$$

Denote the pair of temporal link functions $m := (m_0, m_1)$, and the pair of short-term potential outcome distribution functions by $\gamma := (\gamma_0, \gamma_1) = (P_{S(0)}, P_{S(1)})$. Observe that γ consists of the marginal distributions $P_{S(d)}$, and is not a joint-distribution function of $(S(0), S(1))$.

Is this a good place?Section 3 will reveal that the addition of experimental data brings *no identifying power* without additional modeling assumptions. Modeling assumptions may be classified as: *selection assumptions*, restricting the relationship between $(Y(1), Y(0), S(1), S(0))$ and D; and *treatment response assumptions*, restricting how $(Y(1), Y(0), S(1), S(0))$ are related to each other.

Given functions (m, γ), the corresponding value of τ follows by (7). Relying on this, the approach identifies (m, γ) as an intermediate step towards identifying τ. This has two benefits. First, it will produce sharp bounds for a broad class of modeling assumptions, removing the need for proving sharpness for each assumption. Second, it allows one to account for imperfect compliance in the experiment by permitting partial identification of γ, which is of great practical relevance.

To formalize the class of modeling assumptions, let $\mathcal{M}$ be the set of all link functions, i.e. measurable functions mapping $\mathcal{S}\times\mathcal{S} \to \mathcal{Y}\times\mathcal{Y}$.[7] I assume that the researcher knows or can identify the subset $\mathcal{M}^A \subseteq \mathcal{M}$ to which m belongs, which represents a generic modeling assumption.

Assumption MA. *(Modeling Assumption)* $m \in \mathcal{M}^A \subseteq \mathcal{M}$ *for a known or identified set* $\mathcal{M}^A$.

Assumption MA can accommodate both treatment response and selection assumptions. I will introduce two treatment response assumptions in Section 2.3. **??** explains that Assumption MA nests existing selection assumptions and approaches. Thus, the identification framework will directly extend previously proposed approaches by allowing them to account for imperfect compliance.

Let $\mathcal{H}(\cdot)$ be the sharp identified set for a specified parameter. Finding all (m, γ) consistent with the data and maintained assumptions, including any restriction in the form of Assumption MA, yields $\mathcal{H}(m, \gamma)$. In turn, by the identity (7), $\mathcal{H}(\tau)$ follows directly. To this end, define the

7. More precisely, $\mathcal{M}$ is the set of Borel-measurable functions $\mu : \mathcal{S}\times\mathcal{S} \to \mathcal{Y}\times\mathcal{Y}$ such that $\mu \circ \varsigma$ is P-integrable for some $\mathcal{F}/\mathcal{B}(\mathcal{S}\times\mathcal{S})$-measurable function $\varsigma : \Omega \to \mathcal{S}\times\mathcal{S}$. I refer to random variables m_d and measurable functions μ_d such that $m_d = \mu_d(S(d))$ a.s. as temporal link functions. I also denote $\mu_d(s)$ by $m_d(s)$. This is without loss, since m_d is $\sigma(S(d))$ measurable if and only if $\mu_d : \mathcal{S} \to \mathcal{Y}$ exists by the Doob-Dynkin lemma.

functional $T : \mathcal{M} \times \mathcal{P}^{\mathcal{S}} \times \mathcal{P}^{\mathcal{S}} \to \bar{\mathbb{R}}$, where $\mathcal{P}^{\mathcal{S}}$ collects distribution functions supported on $\mathcal{S}$:

$$T(m, \gamma) = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s). \tag{8}$$

By definition, the sharp identified set $\mathcal{H}(\tau)$ is then equivalent to the set of values T can produce over the sharp identified set $\mathcal{H}(m, \gamma)$:

$$\mathcal{H}(\tau) = \{T(m, \gamma) : (m, \gamma) \in \mathcal{H}(m, \gamma)\}. \tag{9}$$

I first provide $\mathcal{H}(m, \gamma)$ and the corresponding $\mathcal{H}(\tau)$ by the definition. Then, I develop a tractable characterization of the identified set.

2.3 Modeling Assumptions

As Section 3 also shows, plausible inference hinges on plausible modeling assumptions. I thus propose treatment response assumptions that may be defensible based on economic theory or intuition.

Assumption LIV. *(Latent Monotone Instrumental Variable) For any $m \in \mathcal{M}^A$ and $s, s' \in \mathcal{S}$ such that $s < s'$ it holds that $m_d(s) \leq m_d(s')$ for $d \in \{0, 1\}$.*

Assumption LIV has an intuitive interpretation. It posits that the mean of the long-term *potential outcome* $Y(d)$ is non-decreasing conditional on the short-term *potential outcome* $S(d)$. One can symmetrically assume that $E[Y(d)|S(d) = s]$ is non-increasing in s. Results follow directly by defining $\tilde{S}(d) = -S(d)$ and observing that $E[Y(d)|\tilde{S}(d) = s]$ satisfies LIV.

Example 2. *(LIV and Head Start)* In the context of the illustrative application, LIV means that people with higher *potential* childhood test scores $S(d)$ under Head Start participation d, on average, also have weakly higher *potential* earnings in adulthood $Y(d)$ under d.

LIV is related to the monotone instrumental variable (MIV) assumption of Manski and Pepper (2000) (see also Manski and Pepper (2009)). MIV maintains that there exists a variable $V \in \mathcal{V}$ such that $E[Y(d)|V = v]$ is non-decreasing in $v \in \mathcal{V}$, which is observed for *all* individuals. The critical distinction is that the conditioning variable in Assumption LIV is latent. This introduces further complexity, which will be addressed by the identification approach.

Assumption TI. *(Treatment Invariance - TI) For all $m \in \mathcal{M}^A$ and $s \in \mathcal{S}$, $m_1(s) = m_0(s)$.*

The assumption intuitively states that the relationship between the *potential* outcomes $S(d)$ and mean long-term *potential* outcomes $Y(d)$ does not vary with the underlying treatment d.

Example 3. *(TI and Head Start)* Fix a childhood test score $s \in \mathcal{S}$ and suppose we could identify the groups of individuals who would achieve s when subjected to Head Start, and those who would achieve s when excluded from Head Start. TI implies that the two groups would have the same mean earnings under the treatments that induce the test score s.

Another way to clarify the meaning of TI would be through the following separable model:

$$Y(d) = \phi_d(S(d)) + \varepsilon_d = \phi(S(d)) + \varepsilon_d, \;\; \varepsilon_1 \sim \varepsilon_0, \;\; \varepsilon_{d'} \perp\!\!\!\perp S(d), \forall d, d' \in \{0,1\}. \tag{10}$$

The production function ϕ_d and the distributions of unobservables ε_d do not depend on d. It is easy to see that $E[Y(d)|S(d) = s] = \phi(s) + E[\varepsilon]$ which is invariant to d, so TI is implied by the model. Hence, researchers may utilize TI whenever they find such a model to be plausible.

For example, García et al. (2020) argue the plausibility of a similar model when the treatment is an early childhood intervention. They do so using mediation results in Heckman, Pinto, and Savelyev (2013) and extensive falsification testing. They then identify τ by combining observational and experimental data in the special case where $P_O(D = 0) = 1$ and compliance is perfect, i.e. when there is no selection in either dataset. I will show that an implication of their model may be informative of τ even when selection may be present in either dataset.

In the special case of perfect compliance, TI is implied by widely-used statistical surrogacy assumption of Prentice (1989) – $Y \perp\!\!\!\perp S|D, G = E$. Remark 10 in Section A.1 explains the differences.

To connect the findings of this paper with previous results, I will refer to a widely used selection assumption introduced by Athey, Chetty, and Imbens (2020).

Assumption LUC. *(Latent Unconfoundedness) For all* $d \in \{0,1\} : Y(d) \perp\!\!\!\perp D|S(d), G = O$.

According to Chen and Ritzwoller (2023): "Informally, LUC states that all unobserved confounding in the observational sample is mediated through the short-term outcomes". Park and Sasaki (2024a) describe it as a "statistical assumption" and derive equivalent characterizations in a restricted non-parametric model to assign it an economic interpretation.

Plausibility of LUC was brought into question in contexts of economic interest such as early childhood interventions and job-training programs. In the former case, parental interference and the child's inherent ability may be confounding factors for $(D, S(d), Y(d))$, invalidating the assumption. In the latter, the confounding factors may be worker's innate motivation and resourcefulness. For more details see Ghassami et al. (2022) and Imbens et al. (2024). For examples of its use, see Hu, Zhou, and Wu (2022), Park and Sasaki (2024b), Aizer et al. (2024).

Remark 1. Existing approaches are subsumed under Assumption MA. For example, Assumption LUC can be restated as $\mathcal{M}^{LUC} = \{m \in \mathcal{M} : m_d(s) = E_O[Y|S = s, D = d], \forall s \in \mathcal{S}\}$. One can do

the same for the outcome bridge function approach of Imbens et al. (2024, Theorem 1). Let S_t for $t \in \{1,2,3\}$ be subvectors of S. Then under the corresponding assumptions: $\mathcal{M}^{Bridge} = \{m \in \mathcal{M} : m_d(s_3, s_2) = h(s_3, s_2, d),\ h \text{ solves } E_O[Y|S_2, S_1, D] = E_O[h(S_3, S_2, D)|S_2, S_1, D, G = O]\}$.

3 Results on the Roles of Assumptions and Data

To motivate the main results in the next section, I first uncover the roles of the datasets and modeling assumptions. Since $S(d)$ may be observed for some individuals in both datasets, both datasets provide information on their distributions γ. The sole benefit of experimental data, therefore, lies in offering *additional* information on γ. I now examine how this can be beneficial.

Let $\mathcal{H}^O(m, \gamma)$ be the sharp identified set for (m, γ) if only observational data are used. Similarly, let $\mathcal{H}(m, \gamma)$ be the sharp identified set for (m, γ) when both datasets are used. If (m, γ) are consistent with both datasets, they must be consistent with just one dataset under the same assumptions. Thus, $\mathcal{H}(m, \gamma) \subseteq \mathcal{H}^O(m, \gamma)$. Usually $\mathcal{H}(m, \gamma) \subsetneq \mathcal{H}^O(m, \gamma)$ with or without modeling assumptions in the observational data because experimental data provide additional information on γ. By definition (20), the corresponding sharp identified sets for τ are:

$$\begin{aligned} \mathcal{H}^O(\tau) &= \left\{T(m, \gamma) : \ \ (m, \gamma) \in \mathcal{H}^O(m, \gamma)\right\} \\ \mathcal{H}(\tau) &= \left\{T(m, \gamma) : \ \ (m, \gamma) \in \mathcal{H}(m, \gamma)\right\} \end{aligned} \tag{11}$$

recalling that $T(m, \gamma) = \int_{\mathcal{S}} m_1(s) d\gamma_1(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0(s)$. By definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^O(\tau)$. Similarly, let $\mathcal{H}^O(P_{Y(0),Y(1)})$ and $\mathcal{H}(P_{Y(0),Y(1)})$ be the corresponding sharp identified sets for the distribution function $P_{Y(0),Y(1)}$ and observe that $\mathcal{H}(P_{Y(0),Y(1)}) \subseteq \mathcal{H}^O(P_{Y(0),Y(1)})$

Central Role of Modeling Assumptions

I first ask whether it is possible to have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^O(\tau)$ if no modeling assumptions are imposed, which would be desirable. Then, the additional identifying power would solely be the result of random assignment in the experiment. However, this is not the case.

Proposition 1. *Suppose Assumptions RA and EV hold. Then:*

i) $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$*;*

ii) $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$.

See proof on page 48.

On their own, the experimental data bring *no identifying power for* τ or any functional of $P_{Y(0),Y(1)}$. Modeling assumptions in the observational data are *central* in the identification

argument for τ, mirroring their importance in conventional observational studies. They are *necessary* to benefit from the existence of the short-term experiment. Corollary 2 in Appendix A.1 further proves that without such assumptions: 1) τ is unindentified so $\mathcal{H}(\tau) = \mathbb{R}$; 2) $\mathcal{H}(\tau)$ is equivalent to the bounds of Manski (1990) when the support of $Y(d)$ is bounded. Modeling assumptions in observational data are thus necessary to identify at least the sign of τ.

The intuition behind the result is simple. Since $S(d)$ is revealed whenever $D = d$, experimental data only provide more information on the distribution of $S(d)$ for individuals who choose $D \neq d$. However, for them, no data restrict the relationship between $Y(d)$ and $S(d)$. If this relationship is left unrestricted, then additional information on $S(d)$ does not yield more information on $Y(d)$.

Remark 2. A seemingly similar analysis can be found in Park and Sasaki (2024b); however, the conclusion is fundamentally different. First, they consider a different parameter – the treatment effects on treated survivors (ATETS) from Vikström, Ridder, and Weidner (2018). Second, they find that observational data alone yield worst-case bounds on the ATETS, but combined data may be more informative under Assumption LUC. They thus do not uncover the central role of modeling assumptions. I show that observational data are equally informative of any functional of $P_{Y(0),Y(1)}$ as combined data in the absence of modeling assumptions.

Auxiliary Amplifying Role of Experimental Data

Since modeling assumptions are central, experimental data have an *auxiliary role*. To make the role precise, continue to denote by $\mathcal{H}^O(\tau)$ the identified set for τ when only observational data are used and no modeling assumptions are imposed, and let $\mathcal{H}^{O/A}(\tau)$ denote the identified set when a modeling assumption is added. Finally, denote by $\mathcal{H}(\tau)$ the identified set when combined data are used, and the modeling assumption is imposed. It is easy to see that by definition $\mathcal{H}(\tau) \subseteq \mathcal{H}^{O/A}(\tau) \subseteq \mathcal{H}^O(\tau)$.

Proposition 1 shows that without modeling assumptions even point identifying γ does not result in tighter bounds on τ. Thus, any modeling assumption that restricts only γ cannot provide more information on τ. Therefore, any set of assumptions that has identifying power for τ must also restrict m. This yields the following observations.

First, since some information on γ is available from observational data, modeling assumptions restricting m may be informative of τ even in the absence of experimental data. It is possible that $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^O(\tau)$. Second, once modeling assumptions are imposed, more information on γ may further improve the informativeness of the modeling assumptions. In other words, experimental data may *amplify* the identifying power of modeling assumptions, so it may be $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}(\tau)$. Third, $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible. So experimental data do not *necessarily* amplify the identifying power of modeling assumptions. The following remark illustrates these three points using

Assumption LUC. Similar results can be derived for other selection assumptions.

Remark 3. Proposition 3 in Appendix A.1 demonstrates that: 1) LUC provides identifying power for τ without experimental data for common data distributions, so $\mathcal{H}^{O/A}(\tau) \subsetneq \mathcal{H}^{O}(\tau)$ is possible; 2) Since LUC point identifies τ with combined data, usually $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/A}$; 3) there exist data distributions for which LUC point identifies τ without experimental data, so $\mathcal{H}(\tau) = \mathcal{H}^{O/A}(\tau)$ is possible.

Importance of Plausible Modeling Assumptions

In terms of the importance of modeling assumptions, approaches that rely on data combination effectively conduct observational studies. The amplifying role of the experimental data bolsters the importance of plausible modeling assumptions. If the assumptions fail, adding experimental data may be *detrimental*. To see this, suppose a modeling assumption fails and let $\tilde{\mathcal{H}}$ be the misspecified identified set for τ following from combined data. Similarly, let $\tilde{\mathcal{H}}^{O/A}$ be the misspecified set that follows from observational data under the same assumptions. Any value consistent with both datasets must be consistent with just one dataset, so $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$.

Lemma 1. *(Nested Misspecification) Let $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$ be misspecified identified sets for some parameter τ. Let d be the point-to-set distance defined as $d(A,t) := \inf\{\|t-a\| : a \in A\}$ for $A \subseteq \mathbb{R}$ and $t \in \mathbb{R}$. Then:*

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) \leq d(\tilde{\mathcal{H}}, \tau)$$

See proof on page 49.

Lemma 1 states that further reducing the size of any misspecified identified set *necessarily* produces results that are weakly farther away from the truth. Thus, adding experimental data can only move the resulting identified set farther away from the true τ when a modeling assumption fails. In that case, the researcher may only obtain results closer to the ground truth by discarding the available experimental data, and these results may be informative. If the modeling assumption holds, τ is in both identified sets, and adding experimental data cannot produce results farther away from the truth.

Example 4. Suppose $Y, S \in \{0,1\}$ and that we maintain Assumption LUC. Let the DGP be given by **??**. Then $\tau = 0.2$, $\tilde{\mathcal{H}}^{O/A} = [0.15, 0.4]$ and $\tilde{\mathcal{H}} = \{0.35\}$. To be added, the DGP

Example 4 shows that $\tilde{\mathcal{H}}^{O/A}$ can be strictly closer to τ than $\tilde{\mathcal{H}}$ when the modeling assumption fails under non-pathological data-generating processes and standard assumptions. It also demonstrates that adding experimental data may lead the researcher to incorrectly dismiss the true value of τ. We may have $\tau \notin \tilde{\mathcal{H}}$ and $\tau \in \tilde{\mathcal{H}}^{O/A}$, but never the converse.

Remark 4. Lemma 1 is a general misspecification result. It implies that reducing the size of the identified set can never result in the set being closer to the truth. Here, the reduction may happen through the addition of data. More commonly, it is a result of layering additional assumptions.

4 Main Identification Results

I now formalize the two-step identification approach outlined in Section 2.2. Recall that it first finds the identified set $\mathcal{H}(m,\gamma)$ for (m,γ), which then yields the identified set for τ, $\mathcal{H}(\tau)$, by definition. The goal of this section is to construct a characterization of $\mathcal{H}(m,\gamma)$ that will lead to a tractable implementation of $\mathcal{H}(\tau)$. Implementation and consistent estimation are discussed in Section 5.

Notation: Equality of distribution of two random elements or a random element and a law is denoted by $\overset{d}{=}$ (e.g. $Y \overset{d}{=} P_Y$ and $Y \overset{d}{=} Y'$). A, B and K represent sets. $\mathcal{C}(A)$, $\mathcal{B}(A)$ are the families of all closed and Borel subsets of the set A, respectively. $co(A)$ is the closed convex hull of the set A.

Recall that by Assumption MA: $m \in \mathcal{M}^A$ for a known or identified $\mathcal{M}^A$, and that $\mathcal{P}^{\mathcal{S}}$ is the set of all distribution functions supported on $\mathcal{S}$. The main result will show that:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \varsigma_d \overset{d}{=} \gamma_d, \\ \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\} : \ \ um_d(\varsigma_d) \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \text{ a.s.} \end{array} \right\} \tag{12}$$

where $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$ and $P_O(D = d|\varsigma_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)}$ for all $B \in \mathcal{B}(\mathcal{S})$ with $\gamma_d(B) > 0$. This section focuses on the construction of this set. The following technical condition on the probability space will be required for intermediate results.

Assumption PS. *(Probability Space) All random elements are defined on a common non-atomic probability space $(\Omega, \mathcal{F}, P)$. That is, for any $A \in \mathcal{F}$ with positive measure there exists a measurable $B \subset A$ such that $0 < P(B) < P(A)$.*

Beresteanu, Molchanov, and Molinari (2011) indicate that the assumption is not restrictive in many economic settings. For example, it holds if the space admits a continuous random variable. This does not require any of $Y(d)$, $S(d)$, or X to be continuous, only that it is possible to define a continuous random variable on the probability space. In fact, for identification purposes all relevant random variables and vectors can be discrete, continuous, or mixed.

4.1 Construction of $\mathcal{H}(m, \gamma)$

Recall that $m_d(s) = E[Y(d)|S(d) = s]$ and $\gamma_d = P_{S(d)}$. As explained intuitively, for a sharp characterization of (m, γ), one cannot consider m_d and γ_d individually. This is because the latent conditioning variable $S(d)$ simultaneously generates the conditioning $\sigma-$algebra in m_d and determines γ_d. To address the related technical challenges and provide a characterization of $\mathcal{H}(m, \gamma)$ for a generic modeling assumption, I rely on results in random set theory. I introduce the definitions and results necessary to exposit the construction of $\mathcal{H}(m, \gamma)$ in this section. Appendix B.1 provides a more complete but brief overview of the results used in the proofs. I begin with basic definitions specialized to finite-dimensional Euclidean spaces.

Definition 1. A measurable map $\mathbf{R} : \Omega \to \mathcal{C}(\mathbb{R}^d)$ is called a *random (closed) set*.[8]

Definition 2. A random variable $R : \Omega \to \mathbb{R}^d$ such that $R \in \mathbf{R}$ a.s. is called a *(measurable) selection* of $\mathbf{R}$.

Denote the set of all selections of a random set $\mathbf{R}$ by $Sel(\mathbf{R})$ and the set of all integrable selections by $Sel^1(\mathbf{R})$. Whenever the random variable $||\mathbf{R}|| = sup\{||R|| : R \in Sel(\mathbf{R})\}$ is integrable $E[||\mathbf{R}||] < \infty$, then $\mathbf{R}$ is said to be *integrably bounded* and all of its selections are integrable, $Sel^1(\mathbf{R}) = Sel(\mathbf{R})$.

If we could observe $Y(d)$ and $S(d)$ for all individuals and $d \in \{0, 1\}$, then m and γ would be identified by the data. However, $Y(d)$ and $S(d)$ are necessarily only partially observed for at least some d. With this in mind, consider the information about the counterfactuals contained in the observed data. Define the following two closed random sets:

$$\mathbf{Y}(d) = \begin{cases} \{Y\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}, \quad \mathbf{S}(d) = \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases} \tag{13}$$

When the relevant counterfactual is observed, the random set returns only its value. When the potential outcome is unobserved—such as when $D \neq d$ (or $G = E$ for $Y(d)$)—the random set returns the support of the potential outcome, or all possible corresponding values. This reflects the fact that potential outcomes remain unrestricted by the data when unobserved. By definition, $\mathbf{Y}(d)$ and $\mathbf{S}(d)$ summarize all information about $Y(d)$ and $S(d)$ provided by the data.

It is immediate that $Y(d)$ and $S(d)$ are contained in $\mathbf{Y}(d)$ and $\mathbf{S}(d)$ almost surely, i.e. they are their measurable selections. $Y(d)$ and $S(d)$ can be any two selections of the corresponding random sets, but the data do not specify which two. Thus, as Beresteanu, Molchanov, and

8. $\mathbf{R}$ is measurable if for every compact set $K \in \mathcal{K}(\mathbb{R}^d)$: $\{\omega \in \Omega : \mathbf{R}(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$. The codomain $\mathcal{C}(\mathbb{R}^d)$ is equipped by the $\sigma-$algebra generated by the families of sets $\{B \in \mathcal{C}(\mathbb{R}^d) : B \cap K \neq \emptyset\}$ over $K \in \mathcal{K}(\mathbb{R}^d)$.

Molinari (2012) explain, all information in the data about the counterfactuals can be expressed as $S(d) \in Sel(\mathbf{S}(d))$ and $Y(d) \in Sel(\mathbf{Y}(d))$. Then, since $S(d)$ and $Y(d)$ can be any selections of the corresponding random sets, (m, γ) is consistent with the data if and only if there exist selections that "rationalize" them. In other words, (m, γ) are consistent with the data if and only if there exist $\varsigma_d \in Sel(\mathbf{S}(d))$ and $\upsilon_d \in Sel(\mathbf{Y}(d))$ such that $m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d]$ a.s., and $\gamma_d \overset{d}{=} \varsigma_d$ for each $d \in \{0, 1\}$. To then sharply characterize (m, γ), one must consider the additional restrictions on the selections imposed by the assumptions. This is done by Theorem 1.

Theorem 1. *Let Assumptions RA, EV, and MA hold, and* $\tilde{Z} := \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1) \in \tilde{\mathcal{Z}}$. *The sharp identified set for* (m, γ) *is:*

$$\mathcal{H}(m, \gamma) = \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0, 1\},\ \exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z}) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)),\ \varsigma_d \overset{d}{=} \gamma_d,\ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d]\ \ a.s. \end{array} \right\}. \tag{14}$$

where I *is the set of random elements* $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ *such that* $E_1 \perp\!\!\!\perp E_2$.

See proof on page 49.

Theorem 1 yields a direct characterization of the identified set $\mathcal{H}(m, \gamma)$ under any modeling assumption in the form of Assumption MA. This result includes but is not limited to, assumptions and approaches in Section 2.3 and **??**. The theorem removes the need to prove sharpness for each such restriction.

The proof summarizes additional information on the selections $S(d)$ and $Y(d)$ provided by the assumptions, extending the arguments of Beresteanu, Molchanov, and Molinari (2012). The subset of selections $Sel(\mathbf{S}(d))$ where $S(d)$ lies is restricted by Assumptions EV and RA since $S(d)$ may be observed in both datasets. Intuitively, G and Z form a new instrument $\tilde{Z}$, which is independent of $S(d)$ by random assignment of Z and independence of $S(d)$ and G by external validity. Thus, $S(d)$ must be a selection that satisfies this independence condition, which is reflected by $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$. $Y(d)$ is unrestricted by the two assumptions as it is observed only for $G = O$. But, it must be an integrable selection of $\mathbf{Y}(d)$, since $E[|Y(d)|] < \infty$. This is reflected by $\upsilon_d \in Sel^1(\mathbf{Y}(d))$. Finally, m must be in $\mathcal{M}^A$, by the modeling assumption. This representation is intuitive but intractable. To operationalize it, I remove the need to search over the selections $Sel((\mathbf{S}(d), \tilde{Z}))$ and $Sel^1(\mathbf{Y}(d))$. For that, note that for each selection ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z}))$, one can collect all conditional expectations $m_d = E[\upsilon_d|\varsigma_d]$ over $\upsilon_d \in Sel^1(\mathbf{Y}(d))$ into a set. This yields the *random* set $\{E[\upsilon_d|\varsigma_d] :\ \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$. Then, it is easy to see that for a given ς_d:

$$\exists \upsilon_d \in Sel^1(\mathbf{Y}(d)) :\ \ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d] \text{ a.s.} \Leftrightarrow m_d(\varsigma_d) \in \{E[\upsilon_d|\varsigma_d] :\ \ \upsilon_d \in Sel^1(\mathbf{Y}(d))\} \text{ a.s.} \tag{15}$$

In relevant cases this random set is equivalent to a particular random set —the *conditional Aumann expectation*, denoted by $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$.[9] $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ is a unique non-empty random set for each ς_d (Molchanov (2017, Theorem 2.1.71)). Since ς_d is a random vector, when $\mathbf{Y}(d)$ is integrably bounded Li and Ogura (1998, Theorem 1) shows that $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d] = \{E[\upsilon_d|\varsigma_d] : \ \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ a.s., which is exactly the set in condition (89). Hence (89) holds if and only if $m_d(\varsigma_d) \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ almost surely. Properties of the conditional Aumann expectation will then yield an equivalent characterization of the condition which does not include searching over $Sel^1(\mathbf{Y}(d))$.

Theorem 2 combines the properties of conditional Aumann's expectation and Artstein's theorem (Artstein (1983, Theorem 2.1)) to additionally remove the need to search over selections $Sel((\mathbf{S}(d), \tilde{Z}))$. It thus provides an equivalent characterization of $\mathcal{H}(m, \gamma)$ that will lead to a tractable characterization of $\mathcal{H}(\tau)$.

Theorem 2. *Let Assumptions RA, EV, MA, and PS hold. If $\mathbf{Y}(d)$ is integrably bounded, the sharp identified set for (m, γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \varsigma_d \overset{d}{=} \gamma_d, \\ \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\} : \ \ um_d(\varsigma_d) \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \ \ a.s. \end{array} \right\} \tag{16}$$

where $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$ and $P_O(D = d|\varsigma_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)}$ for all $B \in \mathcal{B}(\mathcal{S})$ with $\gamma_d(B) > 0$.

See proof on page 51.

As mentioned, the theorem relies on two groups of results. First, it uses Artstein's theorem to equivalently characterize the restrictions on the distribution γ_d imposed by the condition $\gamma_d \overset{d}{=} \varsigma_d$ for a given $\varsigma_d \in Sel(\mathbf{S}(d)) \cap I$ as:

$$\forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right). \tag{17}$$

Second, it uses the convexification property of the conditional Aumann expectation $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ on non-atomic probability spaces to equivalently characterize $m_d(\varsigma_d) \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ a.s. for a given $\varsigma_d \in Sel(\mathbf{S}(d)) \cap I$. This is done using the support function of the convex hull of the set $\mathbf{Y}(d)$ $h_{co(\mathbf{Y}(d))}(u) := sup_{y \in co(\mathbf{Y}(d))} uy$ (Molchanov (2017, Theorems 2.1.72, 2.1.77), Rockafellar (1970, Theorem 13.1)). Standard arguments in the proof yield that $m_d(\varsigma_d) \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ a.s. if and only

9. The condititonal Aumann expectation is defined with respect to any conditioning sub-σ-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$. Here, this is $\sigma(\varsigma_d)$, which I keep implicit for ease of notation. See Section B.1 for a formal definition.

if:

$$um_d(\varsigma_d) \leq E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s. } \forall u \in \{-1,1\}. \tag{18}$$

Artstein's theorem and the conditional Aumann expectation are commonly used to derive tractable characterizations of identified sets constructed using random sets (for examples see Galichon and Henry (2011), Beresteanu, Molchanov, and Molinari (2011), Beresteanu, Molchanov, and Molinari (2012) and Han and Kaido (2024)). The main technical contribution of Theorem 2 lies in bringing together the two arguments when the relevant conditioning σ-algebra is generated by a *selection* of a random set, rather than an observed random variable, like in Chesher and Rosen (2017). To explain the importance of this, note that standard arguments would lead to the intermediate characterization via expressions (17) and (18):

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \exists \varsigma_d \in Sel(\mathbf{S}(d)) \cap I, \varsigma_d \overset{d}{=} \gamma_d, \\ \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\} : \ \ um_d(\varsigma_d) \leq E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s.} \end{array} \right\} \tag{19}$$

This representation resembles the one in Theorem 2, but differs critically in the need to search over selections $\varsigma_d \in Sel(\mathbf{S}(d)) \cap I$, which would not lead to a tractable characterization of $\mathcal{H}(\tau)$. While expressions (17) and (18) may give the appearance that we may simply search over possible (m,γ) for those that satisfy the two conditions for any random variable such that $\varsigma_d \overset{d}{=} \gamma_d$, this is not the case.

To understand this subtle point, note that by Artstein's theorem, γ_d satisfies (17) if and only if there exists a selection $\varsigma_d \overset{d}{=} \gamma_d$ such that $\varsigma_d \in Sel(\mathbf{S}(d)) \cap I$. It does *not* imply that *every* ς_d such that $\varsigma_d \overset{d}{=} \gamma_d$ is necessarily a selection in $Sel(\mathbf{S}(d)) \cap I$. Thus, not every $\varsigma_d \overset{d}{=} \gamma_d$ needs to generate the same conditioning σ-algebra.[10] This σ-algebra determines the restrictions imposed on m_d via $E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$. Hence, an arbitrary $\varsigma_d \overset{d}{=} \gamma_d$ may not yield the appropriate restrictions on m_d. For this reason, one would need to search through $\varsigma_d \in Sel(\mathbf{S}(d)) \cap I$ to construct the identified set.

Theorem 2 addresses this issue. It equivalently restates the condition $m_d(\varsigma_d) \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ a.s. so that it becomes invariant to the conditioning random variable ς_d up to its distribution $\varsigma_d \overset{d}{=} \gamma_d$. In this representation, any two $\varsigma'_d \overset{d}{=} \varsigma_d \overset{d}{=} \gamma_d$ will lead to identical restrictions on m_d imposed by the data. This is crucial for producing a tractable implementation of $\mathcal{H}(\tau)$ in the

10. For example, consider the standard probability space $([0,1], \mathcal{B}([0,1]), \lambda)$. Let $R = \mathbb{1}[\omega \geq \frac{2}{3}]$ and $R' = \mathbb{1}[\omega \leq \frac{1}{3}]$. Then $R \overset{d}{=} R'$, but $\sigma(R) \neq \sigma(R')$.

next section. Before that, recall that results discussed above yield the sharp identified set $\mathcal{H}(\tau)$, by definition:

$$\mathcal{H}(\tau) = \{T(m,\gamma) : (m,\gamma) \in \mathcal{H}(m,\gamma)\} \tag{20}$$

where $T(m,\gamma) = \int_{\mathcal{S}} m_1(s)d\gamma_1(s) - \int_{\mathcal{S}} m_0(s)d\gamma_0(s)$.

5 Implementation and Estimation

The previous section provides an abstract representation of the identified set $\mathcal{H}(\tau)$. This section shows that $\mathcal{H}(\tau)$ may be restated as an interval bounded by two generalized bilinear optimization problems in commonly considered settings. Using this, it then develops a consistent estimator for $\mathcal{H}(\tau)$.

5.1 Characterization of $\mathcal{H}(\tau)$

I maintain that $\mathcal{Z}$ is a bounded set.

Nonparametric treatment response assumptions typically deliver identifying power for means under support restrictions. While such restrictions are not always necessary (see Remark 5 and Manski (1997), Manski and Pepper (2000)), they are usually needed for informative identified sets on means when only treatment response assumptions are maintained.

Assumption BS. *(Bounded Support)* $P(Y(d) \in [Y_L, Y_U]) = 1$ *for some known finite* Y_L, Y_U.

Assumption BS is standard in the partial identification literature. It states that the support of $Y(d)$ is contained by some known closed interval. I normalize the interval to $[0,1]$ without loss of generality. This assumption may be natural for various $Y(d)$ such as binary indicators, or discrete and continuous variables that are logically bounded. For some $Y(d)$, it may be restrictive.

Remark 5. *(Novel Treatment)* When $P_O(D = d) = 0$ for some $d \in \{0,1\}$, Assumption TI point identifies $E[Y(d')|S(d') = s]$ for $d' \in \{0,1\}$ and $s \in \mathcal{S}$ (up to almost sure equivalence). Then τ will be point identified without bounded support of $Y(d)$ if compliance is perfect, because the experiment identifies distributions of $S(d')$. If compliance is imperfect, one may still obtain informative bounds for τ without support restrictions.

start with bounded support assumptions here

Identified sets following from Artstein's theorem generally cannot be fully characterized when relevant outcome spaces are infinite, even if the observed data distributions are known. This can

also be easily seen here. To verify if a candidate (m, γ) is in the identified set, one must establish that each γ_d satisfies an inequality condition for each closed subset $B \in \mathcal{C}(\mathcal{S})$. If $\mathcal{S}$ is infinite, then so is $\mathcal{C}(\mathcal{S})$.[11] A common way of addressing this issue is to discretize the relevant variables or focus on settings where they are finitely-supported (Galichon and Henry (2011), Russell (2021), Ponomarev (2024)). I do the same.

Henceforth I maintain that $S(d) \in \mathcal{S} = \{1, 2, \ldots, k\}$, either by definition or following discretization performed by the researcher. Subtleties related to the interpretation of results under discretization are discussed in Section A.4. I do not require the long-term outcome support $\mathcal{Y}$ to be a finite or discrete set. I maintain Assumption BS for simplicity since the focus is on treatment response assumptions which usually require bounded support to be informative. One can represent γ_d as an element of a $k-$dimensional simplex $\Delta(k)$, and $\gamma \in \Delta(k) \times \Delta(k)$. Similarly, $m \in \mathcal{M} = \mathcal{Y}^k \times \mathcal{Y}^k$, and the modeling assumption can be represented as $\mathcal{M}^A \subseteq \mathcal{Y}^k \times \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the $s-$th element of the corresponding vectors. This leads to the following characterization result.

Theorem 3. *Let Assumptions RA, EV, MA, PS and BS hold. Suppose $\mathcal{S}$ is a finite set and that $\mathcal{M}^A$ is closed and convex. Then:*

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right].$$

See proof on page 54.

By the theorem, $\mathcal{H}(\tau)$ can be equivalently restated as an interval bounded by solutions to two optimization problems where $\mathcal{H}(m, \gamma)$ represents the constraint set. This characterization follows for easily verifiable high-level conditions on $\mathcal{M}^A$. Remark 6 explains that these conditions are satisfied by proposed and existing assumptions.

Using optimization problems to characterize identified sets has become common in partial identification analyses (for recent examples, see Mogstad, Santos, and Torgovitsky (2018), Torgovitsky (2019), Russell (2021), Kamat (2024) and references therein). Such representations usually follow directly from the convexity of the constraint set and the continuity of the objective function. Here, a slightly more involved argument is required since T is a difference of two Riemann-Stieltjes integrals, thus bilinear and hence only separately continuous in m and γ. The proof shows that T is jointly continuous and that $\mathcal{H}(m, \gamma)$ is a compact and convex set under the assumptions of the theorem. Then $\mathcal{H}(\tau)$ is a continuous image of a compact and convex set by definition, hence a compact and connected set, i.e., a closed interval.

11. One may hope to alleviate the issue by using a *core-determining class*, i.e. a subfamily of $\mathcal{C}(\mathcal{S})$ sufficient to summarize all restrictions on γ (Galichon and Henry (2011)). However, even the smallest core-determining class will contain infinitely many sets (Ponomarev (2024, Theorem 1)).

Remark 6. Assumptions LIV and TI lead to $\mathcal{M}^A$ that is closed and convex when $|\mathcal{S}| < \infty$. Moreover, whenever m is identified by the data, such as under LUC, $\mathcal{M}^A$ is a singleton and hence closed and convex.

5.1.1 Reducing Computational Complexity

The distinguishing feature of the setting in this paper is that optimization problems in Theorem 3 represent *generalized bilinear programs* (see Al-Khayyal (1992)). Such programs are computationally demanding in general. I thus propose further simplifications that exploit the structure of the identified set $\mathcal{H}(m, \gamma)$. I first restate the problems as bilevel programs and show that inner problems have closed-form solutions under proposed assumptions.[12] I then utilize the concept of core-determining classes to further reduce the number of constraints imposed by $\mathcal{H}(m, \gamma)$.

First, decompose $\mathcal{H}(m, \gamma)$ into its projection $\mathcal{H}(\gamma) := \{\gamma' : \exists m' \text{ s.t. } (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ and corresponding fibers $\mathcal{H}(m|\gamma') := \{m' : (m', \gamma') \in \mathcal{H}(m, \gamma)\}$ at each $\gamma' \in \mathcal{H}(\gamma)$. The fibers form a correspondence $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$. The identified set can then be restated as:

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \min_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} \max_{\tilde{m} \in \mathcal{H}(\tilde{m}|\tilde{\gamma})} T(\tilde{m}, \tilde{\gamma}) \right]. \tag{21}$$

The inner optimization problems may have known closed-form solutions given by some selectors L_γ and U_γ of the correspondence $\mathcal{H}(m|\cdot)$. This is formalized by the following definition.

Definition 3. (Minimal and Maximal Selectors) Let $\mathcal{H}(m|\cdot) : \mathcal{H}(\gamma) \rightrightarrows \mathcal{M}^A$ be a correspondence defined by fibers $\mathcal{H}(m, \gamma)$ over its projection $\mathcal{H}(\gamma)$. L_γ is a *minimal selector with respect to* T if for any $\gamma \in \mathcal{H}(\gamma)$: $T(L_\gamma, \gamma) \leq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$. U_γ is a *maximal selector with respect to* T if for any $\gamma \in \mathcal{H}(\gamma)$: $T(U_\gamma, \gamma) \geq T(m, \gamma)$ for all $m \in \mathcal{H}(m|\gamma)$.

Corollary 1. *Let conditions of Theorem 3 hold. If $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to T, then:*

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma}), \max_{\tilde{\gamma} \in \mathcal{H}(\gamma)} T(U_{\tilde{\gamma}}, \tilde{\gamma}) \right].$$

See proof on page 57.

Remark 7 explains that this simplification is possible for individual modeling assumptions I propose, as well as existing assumptions from the literature.

Remark 7. (Assumptions with Minimal/Maximal Selectors) Lemma 11 shows that Assumptions LIV and TI produce known minimal and maximal selectors. Moreover, whenever m is identified, such as under LUC, minimal and maximal selectors exist and coincide by definition.

12. An unrelated example of using bilevel optimization problems for identification can be found in Moon (2024).

Corollary 1 provides a computationally appealing formulation of the identified set. However, note that even when the problems are simplified via minimal/maximal selectors, the number of constraints imposed by $\mathcal{H}(\gamma)$ in the outer optimization step may be very large. Namely, for each γ_d, there are 2^{k-1} nontrivial sets in $\mathcal{C}(\mathcal{S})$, each of which translates to a linear inequality constraint. The following proposition further reduces the number of constraints on each γ_d to $k-1$.

Proposition 2. *Let Assumptions RA, EV, MA, PS and BS hold. Suppose $\mathcal{S}$ is a finite set. Then the sharp identified set for (m,γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z(P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ \forall u \in \{-1,1\}: \quad u m_d(s) \leq u E[Y|S,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right) \end{array} \right\}. \tag{22}$$

See proof on page 58.

Proposition 2 provides an equivalent representation for $\mathcal{H}(m,\gamma)$ when $|\mathcal{S}| < \infty$. It may be compared to the one characterized by Theorem 2 in Section 4. They differ only in the number of conditions imposed on γ_d for $d \in \{0,1\}$. The theorem imposes an inequality condition for each closed subset $B \in \mathcal{C}(\mathcal{S})$, and the proposition replaces this with and inequality condition over of all singleton subsets $\{\{s\} : s \in \mathcal{S}\}$. This reduces the cardinality of the constraint set for γ_d from 2^{k-1} to $k-1$.

The proposition may thus substantially reduce the number of inequality restrictions on γ_d for both $d \in \{0,1\}$ without any loss of information. This is done by utilizing the concept of a *core-determining* class (Galichon and Henry (2011)). For a formal definition, see Section B.1. Intuitively, a core determining class removes redundant restrictions on γ_d imposed by the formulation of Theorem 2.

Table 1 depicts the magnitude of this reduction, providing the total number of constraints on γ imposed by observed data in a single optimization problem with respect to $|\mathcal{S}|$.[13] Even for relatively few support points, the reduction in the number of constraints is sizeable. If $S(d)$ represents percentiles, then Theorem 2 yields a prohibitively complex constraint set, while that of the proposition remains manageable.

Remark 8. Optimization problems $\max / \min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})$ become linear and simple to solve in special cases. This happens whenever either $\mathcal{H}(m,\gamma) = \{m\} \times \mathcal{H}(\gamma)$; or $\mathcal{H}(m,\gamma) =$

13. The number of constraints on m imposed by the data is $2k$ and may be affected by modeling assumptions. It is irrelevant when there are minimal/maximal selectors, since then the optimization problem over m is removed.

	$\lvert\mathcal{S}\rvert$				
Constraint # for γ	2	5	10	20	100
Theorem 2	4	32	1024	1048576	$> 10^{30}$
Proposition 2	2	8	18	38	198

Table 1: Number of constraints on γ in $\mathcal{H}(m,\gamma)$.

$\mathcal{H}(m) \times \{\gamma\}$ and $\mathcal{M}^A$ can be expressed using linear constraints. Assumptions that point identify m independently of γ, such as Assumption LUC, yield $\mathcal{H}(m,\gamma) = \{m\} \times \mathcal{H}(\gamma)$. The latter case would occur for Assumptions LIV and TI under perfect compliance. Note that then Lemma 11 would yield a closed-form expression for $\mathcal{H}(\tau) = [T(L_\gamma,\gamma), T(U_\gamma,\gamma)]$ where γ takes a single value.

To do: 1) Lemma 11; Add some simulations or something?

5.2 Estimation

The analysis thus far has assumed knowledge of observational and experimental data distributions. I now develop a consistent estimator of $\mathcal{H}(\tau)$ using finite sample data based on the implementation strategy. Following standard practice, I focus on consistency in terms of the Hausdorff distance defined for sets A and B on Euclidean spaces as $d_H := \max\{\sup_{a\in A}\inf_{b\in B}||a-b||, \sup_{b\in B}\inf_{a\in A}||a-$

Suppose that the researcher observes experimental and observational samples $\{(S_j, D_j, Z_j)\}_{j=1}^{n_E}$ and $\{(Y_i, S_i, D_i)\}_{i=1}^{n_O}$, respectively. Let $n = \min\{n_O, n_E\}$, and denote by $P_{E,n}(S \in A, D = d, Z = z) = \frac{\sum_{j=1}^{n_E} \mathbb{1}\{S_j \in A, D_j = d, Z_j = z\}}{n_E}$ and $P_{O,n}(S \in A, D = d) = \frac{1}{n_O}\sum_{i=1}^{n_O} \mathbb{1}\{S_i \in A, D_i = d\}$ be standard empirical measures. Denote by $E_{E,n}$ and $E_{O,n}$ the corresponding empirical expectations. It is well known that these represent consistent estimators for their population counterparts as $n \to \infty$. Moreover, we note that their population counterparts, along with $\mathcal{M}^A$, fully characterize $\mathcal{H}(m,\gamma)$ and thus $\mathcal{H}(\tau)$. Define the empirical version of $\mathcal{H}(m,\gamma)$:

$$\mathcal{H}_n(m,\gamma) := \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(\sup_{z\in\mathcal{Z}} P_{E,n}(S=s, D=d|Z=z), P_{O,n}(S=s, D=d)\right), \\ m_d(s) \geq E_{O,n}[Y|S=s, D=d]\frac{P_{O,n}(S=s,D=d)}{\gamma_d(s)}, \\ m_d(s) \leq E_{O,n}[Y|S=s, D=d]\frac{P_{O,n}(S=s,D=d)}{\gamma_d(s)} + 1 - \frac{P_{O,n}(S=s,D=d)}{\gamma_d(s)} \end{array} \right\} \quad (23)$$

and let $\mathcal{H}_n(\gamma)$ and $\mathcal{H}_n(m|\gamma)$ represent the projection and fibers of $\mathcal{H}_n(m,\gamma)$, respectively. I propose to estimate $\mathcal{H}(\tau)$ using:

$$\mathcal{H}_n(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} \min_{\tilde{m}\in\mathcal{H}_n(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}_n(\gamma)} \max_{\tilde{m}\in\mathcal{H}_n(\tilde{m}|\tilde{\gamma})} T(m,\tilde{\gamma})\right]. \quad (24)$$

Similarly, if maximal and minimal selectors of $\mathcal{H}(m|\cdot)$ exist, let $U_{n,\gamma}$ and $L_{n,\gamma}$ be their plug-in

sample analogs. Then note that $\mathcal{H}_n(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(U_{\tilde{\gamma}}, \tilde{\gamma})\right]$.

To prove the consistency of $\mathcal{H}_n(\tau)$, I maintain the following regularity condition.

Assumption E. *(Estimation)*

i) $\{(S_j, D_j, Z_j)\}_{j=1}^{n_E}$ *and* $\{(Y_i, S_i, D_i)\}_{i=1}^{n_O}$ *are i.i.d. samples;*

ii) $\mathcal{M}^A$ *can be represented through linear equality and constraints. The Jacobian of linear equality constraints has full row rank.*

iii) $|\mathcal{S}|, |\mathcal{Z}| < \infty$*;*

iv) $int(\mathcal{H}(m, \gamma)) \neq \emptyset$*.*

Comment on interior and assumptions

Theorem 4. *Let Assumptions RA, EV, MA, PS and E hold. Then* $\mathcal{H}_n(\tau) \xrightarrow{p} \mathcal{H}(\tau)$ *in the Hausdorff metric as* $n \to \infty$*.*

The proof adapts the arguments of Russell (2021, Theorem 2). It uses continuity It consists of .... using Shi and Shum (2015)...

Proof. $\mathcal{H}(\tau)$ and $\mathcal{H}_n(\tau)$ are both intervals. To prove $\mathcal{H}_n(\tau) \xrightarrow{p} \mathcal{H}(\tau)$ in the Hausdorff metric, I it is sufficient to show that boundaries of $\mathcal{H}_n(\tau)$ converge in probability to boundaries of $\mathcal{H}_(\tau)$ as $n \to \infty$. I show that for any $\varepsilon > 0$:

$$P\left(\left|\max_{\tilde{m},\tilde{\gamma}\in\mathcal{H}_n(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) - \max_{\tilde{m},\tilde{\gamma}\in\mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}\right| > \varepsilon\right) = 0 \tag{25}$$

and the argument for the lower bound is symmetric. □

Remark 9. Note that assumption LIV and TI can be represented using linear constraints.

5.3 Inference

6 Application

TBA

7 Conclusion

TBA

Appendices

Appendix A Extensions

A.1 Additional Results on the Roles of Data and Assumptions

This appendix collects complementary results for the discussion in Section 3. Suppose first that no modeling assumptions are maintained.

Corollary 2. *Suppose Assumptions RA and EV hold. If $\mathcal{Y} = \mathbb{R}$, the sharp identified set for τ is $\mathcal{H}(\tau) = \mathbb{R}$. If we additionally maintain $\mathcal{Y} = [0, 1]$:*

$$\mathcal{H}(\tau) = [E_O[YD] - E_O[Y(1-D)] - P_O(D=0), E_O[YD] - E_O[Y(1-D)] + P_O(D=1)]\,. \quad (26)$$

In both cases, $0 \in \mathcal{H}(\tau)$ and the sign of τ not identified.

See proof on page 59.

Corollary 2 bears a relation to existing findings. First, it reproduces bounds of Manski (1990), which utilize only the observational dataset. The bounds remain sharp even when the experimental dataset is added since it brings no identifying power, *on its own*. In other words, Assumptions RA and EV have no identifying power, *on their own*.

In the setting of Corollary 2, Athey, Chetty, and Imbens (2020) state "*...maintained assumptions (author's note: RA with perfect compliance and EV) are in general not sufficient point-identification of the average effect of interest. Of course, this does not mean that these assumptions do not have any identifying power. They do, in fact, affect the identified sets in the spirit of the work by (Manski (1990)).*" The corollary shows that bounds of Manski (1990) remain sharp when Assumptions RA and EV are introduced. In that sense, the finding clarifies that the two assumptions do not affect the identified sets, *on their own*. However, they may have an effect in the presence of modeling assumptions.

Athey et al. (2024, Lemmas 1 and 2) provide bounds on long-term treatment effects in a different setting where D is unobserved in the observational data and experimental compliance

is perfect.[14] Their bounds may be narrower than those in Corollary 2, and do not maintain explicit modeling assumptions involving counterfactuals. However, this does not contradict our result. Namely, their bounds are derived under assumptions imposed on outcome variables: 1) $Y \perp\!\!\!\perp D|S, G = E$ (statistical surrogacy - Prentice (1989)); 2) $G \perp\!\!\!\perp Y|S$ (comparability). Lemma 10 and Remark 10 explain that these assumptions on outcomes imply underlying selection assumptions.

Recall that $\mathcal{H}^O(\tau)$ the identified set for τ when only observational data are used and no modeling assumptions are imposed, and let $\mathcal{H}^{O/LUC}(\tau)$ denote the identified set under Assumption LUC. Finally, let $\mathcal{H}(\tau)$ be the identified set when combined data are used under Assumption LUC.

Proposition 3. *Let Assumptions EV and LUC hold.*

i) Suppose the observed data distribution $P_O(Y, S, D)$ is such that $V_O[Y|S, D = d] > 0$ $P-a.s.$ for some $d \in \{0, 1\}$ and that $\mathcal{Y}$ is a bounded set. Then $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

ii) If the observed data distribution $P_O(Y, S, D)$ is such that $E_O[Y|S, D = d]$ is a trivial measurable function for all $d \in \{0, 1\}$, then τ is point-identified, and $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$.

See proof on page 60.

A few observations are in order. First, the proposition shows that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$ is possible. That is, LUC may have identifying power for τ for a large class of observable distributions $P_O(Y, S, D)$ even when experimental data are not used. A sufficient condition for this is that Y is bounded, and that S is not a perfect predictor of Y for at least some $D = d$.

Second, Athey, Chetty, and Imbens (2020) show that $\mathcal{H}(\tau)$ is a singleton under combined data and LUC. Since $\mathcal{H}^{O/LUC}(\tau)$ is not generally a singleton, we usually have $\mathcal{H}(\tau) \subsetneq \mathcal{H}^{O/LUC}$. Consequently, experimental data may *amplify* the identifying power of LUC.

Third, the proposition shows that $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$ is possible. That is, short-term experimental data are not necessary for point identification of τ under LUC. Thus, experimental data do not *necessarily* amplify the identifying power of LUC. This intuitively happens when the short-run outcomes S are not predictive of the mean long-term outcomes Y.[15] This condition is strong and may lack practical applicability. However, the result has important theoretical implications in clarifying the role of the experimental data.

14. More precisely, they bound $E_E[Y(1) - Y(0)]$. These bounds remain valid for τ when Assumption EV is imposed.

15. Observe that no restrictions on $\mathcal{Y}$ are required in this case.

A.1.1 More on Treatment Invariance and Surrogacy

TI is implied by surrogacy when compliance is perfect. One may thus wish to intuitively interpret TI as stating that the treatment effect on the long-term outcome is fully mediated by the short-term outcome, an interpretation commonly used for the surrogacy assumption. However, the remark also explains that surrogacy implies selection assumptions whenever compliance is imperfect. On the other hand, TI is always a treatment response assumption. Moreover, surrogacy is often paired with other assumptions, and they jointly imply selection assumptions even if compliance is perfect. See Remark 10 for details. I make no such restrictions.

Remark 10. By Lemma 10 *ii)*, TI is implied by surrogacy when the experiment features perfect compliance. When compliance is imperfect, surrogacy implies $E_E[Y(1)|S(1) = s, D = 1] = E_E[Y(0)|S(0) = s, D = 0]$ for $s \in \mathcal{S}$, which is an a priori restriction on the selection mechanism of experimental individuals, because $Y(d)$ are never observed for $G = E$

Work relying on surrogacy for identification, such as Athey et al. (2024), commonly also maintains – $G \perp\!\!\!\perp Y|S$ (comparability). Comparability and surrogacy jointly imply a selection assumption even if compliance is perfect. Note that for any $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$\begin{aligned} E[Y(d)|S(d) = s] &= E_O[Y(d)|S(d), D = d]P_O(D = d|S(d) = s) \\ &\quad + E_O[Y(d)|S(d), D \neq d]P_O(D \neq d|S(d) = s) \\ &= E_O[Y(1)|S(1) = s, D = 1]P_O(D = 1|S = s) \\ &\quad + E_O[Y(0)|S(0) = s, D = 0]P_O(D = 0|S = s) \end{aligned}$$

where the first equality is by LIE and the second by Lemma 10 *vi)*. Therefore, for any s and d such that $P(D \neq d, S(d) = s) > 0$ by rearranging terms:

$$\begin{aligned} &E_O[Y(d)|S(d) = s, D \neq d] = \\ &= \frac{E_O[Y|S, D = d]\left(P_O(D = d|S = s) - P_O(D = d|S(d) = s)\right) + E_O[Y|S, D \neq d]P_O(D \neq d|S = s)}{P_O(D \neq d|S(d) = s)} \end{aligned}$$

which relates $(Y(1), Y(0), S(1), S(0))$ and D in the observational data.

A.2 Differing Population Estimands

TO BE ADDED

The LTE, defined as $E[Y(1) - Y(0)|G = g]$, may differ with $G = g$ up to weighting of $\tau(x)$ by the distributions of X. $\tau(x)$ may be used to identify the LTE in any of the two datasets as

$E[\tau(X)|G=g]$. This is discussed further in Appendix A.2.

- $P_g(X)$ can differ with respect to g.
- So $\tau_g = E_g[\tau(X)]$ can differ with respect to g even under Assumption EV.
- This is easily resolved by integration.
- Identification of $\tau(x)$ is sufficient to identify the unconditional average treatment effect (ATE) in any of the two populations as $E[\tau(X)|G=g]$ since the corresponding distributions of X are identified, if $Supp(X|G=O) = Supp(X|G=E)$.

A.3 Testable Implications of External Validity

The corresponding falsification test is the topic of ongoing work (Obradović (2024)). It is not verifiable as that would require showing that $P(Y(1), Y(0), S(1), S(0)|X, G=g)$ is invariant to g. This is precluded because Y is never observed in the experiment (i.e., for $g=E$). However, the assumption is refutable, as it has testable implications. In Section A.3, I present the strongest testable implication. The corresponding falsification test is the topic of ongoing work (Obradović (2024)).

See above copied

External validity of the experiment is a key identifying assumption in the data combination setting. Hence, it is imperative to consider its plausibility. One should argue why the assumption is credible based on the contextual knowledge. However, it would be very appealing if one could also formally test it. So far, no such test procedure has been proposed.

In this section, we first explicitly analyze testability Assumption EV in its unconditional form. Then we extend the test to the case when the assumption is assumed to hold conditionally on X. Consider $G \perp\!\!\!\perp (Y(1), Y(0), S(1), S(0))$ or by definition:

$$P((Y(1), Y(0), S(1), S(0)) \in B|G=O) = P((Y(1), Y(0), S(1), S(0)) \in B|G=E)$$

for any Borel set $B \in \mathcal{B}(\mathcal{Y}^2 \times \mathcal{S}^2)$. Equivalently, we can write the unconditional Assumption EV as:

$$P((Y(1), Y(0), S(1), S(0)) \in B|G=O) = P((Y(1), Y(0), S(1), S(0)) \in B|G=E) \tag{27}$$

for any closed set $B \in \mathcal{C}(\mathcal{Y}^2 \times \mathcal{S}^2)$. Necessity is immediate since any closed set is a Borel set. Sufficiency follows by Dynkin's $\pi - \lambda$ theorem, because closed sets $\mathcal{C}(\mathcal{Y}^2 \times \mathcal{S}^2)$ form a π-system and generate the Borel σ-algebra $\mathcal{B}(\mathcal{Y}^2 \times \mathcal{S}^2)$. (Billingsley (1995, Theorems 3.2, 3.3))

The equality (27) is fundamentally untestable. First, $Y(d)$ is never observed for $G = E$. This leads us to consider tests that only involve distributions $P(S(1), S(0)|G = g)$. Second $(S(1), S(0))$ are never observed jointly, so we will have to focus on marginal restrictions $P(S(d)|G = g)$. This will allow us to construct a falsification test for Assumption EV. An implication of (27) that we consider is:

$$\begin{aligned} P(S(1) \in B|G = O) &= P(S(1) \in B|G = E) \\ P(S(0) \in B|G = O) &= P(S(0) \in B|G = E) \end{aligned} \tag{28}$$

for any closed set $B \in \mathcal{C}(\mathcal{S})$. We can make this set of implications viable for testing by observing that $P(S(d) \in B|G = E)$ and $P(S(d) \in B|G = O)$ can be identified or bounded in experimental and observational populations, respectively. Results in random set theory yield appealing characterizations of relevant identified sets that will facilitate creation of a test. They will allow us to provide the strongest testable implications in the sense that will be made precise below.

To understand how the observed data relate to (28), first focus on $P(S(d)|G = O)$. It is generally partially identified due to selection when $G = O$.[16] Its sharp identified set by Lemma 7 is:

$$\Gamma^O(S(d)) = \{\gamma \in \mathcal{P}^{\mathcal{S}} : \gamma(B) \geq P(S \in B, D = d|G = O) \ \forall B \in \mathcal{C}(\mathcal{S})\}. \tag{29}$$

Equation (29) follows from a direct application of Artstein's inequalities to an appropriately defined random set that collects all observable data. The equation states that a distribution $P(S(d)|G = O)$ is consistent with the data if and only if for every closed subset $B \subseteq \mathcal{S}$ it is true that:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|G = O) \geq P(S \in B, D = d|G = O). \tag{30}$$

Consequently, $P(S(d) \in B|G = O) = P(S(d) \in B|G = E)$ for any closed subset B is consistent with data if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|G = E) \geq P(S \in B, D = d|G = O). \tag{31}$$

Next, we turn to $P(S(d) \in B|G = E)$. Consider first the case of perfect compliance. When

16. More precisely, it must be at most partially identified for at least one $d \in \{0, 1\}$. If and only if $P_O(D = d) = 1$ for some d, then $P_O(S(d))$ is point-identified and $P(S((d')))$ is unindentified for $d \neq d'$.

$Z = D|G = E$ P-a.s., $P(S(d) \in B|G = E) = P(S \in B|D = d, G = E)$ so the distribution of $S(d)$ in experimental data is identified. Then (31) can be rewritten as:

$$\begin{aligned} &\forall B \in \mathcal{C}(\mathcal{S}): \quad P(S \in B|D = 1, G = E) \geq P(S \in B, D = 1|G = O) \\ &\forall B \in \mathcal{C}(\mathcal{S}): \quad P(S \in B|D = 0, G = E) \geq P(S \in B, D = 0|G = O). \end{aligned} \tag{32}$$

If Z may differ from D with positive probability in the experiment so that there is imperfect compliance, then generally we may have $P(S(d) \in B|G = E) \neq P(S \in B|D = d, G = E)$ for some closed set B. In this case, Lemma 7 provides the sharp identified set for $P(S(d)|G = E)$:

$$\Gamma^E(S(d)) = \{\gamma \in \mathcal{P}^{\mathcal{S}} : ess\sup_Z P(S \in B, D = d|Z, G = E)\ \forall B \in \mathcal{C}(\mathcal{S})\}. \tag{33}$$

The characterization states that a distribution $P(S(d)|G = E)$ is consistent with the data if and only if for every closed subset $B \subseteq \mathcal{S}$ it is true that:

$$\forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|G = E) \geq ess\sup_Z P(S \in B, D = d|Z, G = E). \tag{34}$$

Recall, that (30) yields the sharp identified set for distributions $P(S(d)|G = O)$ that are consistent with the data. Then if no distributions that satisfy (34) are also consistent with (30), we know that $P(S(d)|G = E) \neq P(S(d)|G = O)$. Hence, the testable implications can be stated as:

$$\begin{aligned} &\exists \gamma_1 \in \Gamma^E(S(1)) \text{ s.t. } \forall B \in \mathcal{C}(\mathcal{S}): \quad \gamma_1(B) \geq P(S \in B, D = 1|G = O) \\ &\exists \gamma_0 \in \Gamma^E(S(0)) \text{ s.t. } \forall B \in \mathcal{C}(\mathcal{S}): \quad \gamma_0(B) \geq P(S \in B, D = 0|G = O). \end{aligned} \tag{35}$$

Observe that when there is perfect compliance, $\Gamma^E(S(d)) = \{P(S|D = d, G = E)\}$ so that (35) simplifies to (32).

The formulation of the implications lays bare the necessity of implications and lack of sufficiency for Assumption EV. Namely, if the implications hold, that does not mean that $P(S(d)|G = E) = P(S(d)|G = O)$, only that we cannot deduce that $P(S(d)|G = E) \neq P(S(d)|G = O)$. Hence, if the implications fail, we can conclude that Assumption EV does not hold. But if they hold, that does not imply that Assumption EV is valid. Thus, we can only refute the assumption, but never verify it. This limitation is inherent to various specification tests such as classical overidentification testing in the homogenous effect linear instrumental variable setting, and testing of LATE restrictions of Imbens and Angrist (1994). (Kitagawa (2015), Huber, Laffers, and Mellace (2015), Mourifié and Wan (2017))

Since these conditions are only necessary, Proposition 4 explores whether there exist stronger conditions that are testable.

Proposition 4. *Suppose Assumption RA holds. If implications (35) hold, then there exist conditional distributions* $\tilde{P}(Y(1), Y(0), S(1), S(0), D, Z|G = E)$ *and* $\tilde{P}(Y(1), Y(0), S(1), S(0), D|G = O)$ *that generate observed data such that Assumption EV holds.*

Proof. The proof is constructive. For each $G = g$ I first construct the marginals and then produce the relevant joint distribution using the trivial (independent) coupling.

Fix $\gamma_d \in \Gamma_d^E(S(d))$ for $d \in \{0, 1\}$ such that (35) hold. Then for $d \in \{0, 1\}$ by Lemma 7:

$$\begin{aligned} &\forall B \in \mathcal{C}(\mathcal{S}): \quad \gamma_d(B) \geq P(S \in B, D = d|G = O) \\ &\forall B \in \mathcal{C}(\mathcal{S}): \quad \gamma_d(B) \geq ess\sup_Z P(S \in B_S, D = d|Z, G = E) \end{aligned} \tag{36}$$

By Dynkin's $\pi - \lambda$ theorem, we can replace closed sets $B \in \mathcal{C}(\mathcal{S})$ with all Borel sets $B \in \mathcal{B}(\mathcal{S})$. Let $B_Z \in \mathcal{B}(\mathcal{Z})$, $B_Y \in \mathcal{B}(\mathcal{Y})$ and $B_S \in \mathcal{B}(\mathcal{S})$ be generic sets in their respective Borel $\sigma-$algebras.

Observe that for $\tilde{P}(Y(1), Y(0), S(1), S(0), D, Z|G = E)$ and $\tilde{P}(Y(1), Y(0), S(1), S(0), D|G = O)$ to be consistent with observed data $P(S, D, Z|G = E)$ and $P_O(Y, S, D)$, we must have:

- $\tilde{P}(D = d|G = O) = P(D = d|G = O)$;
- $\tilde{P}(D = d, Z \in B_Z|G = E) = P(D = d, Z \in B_Z|G = E)$.

We then need to construct conditional distributions $\tilde{P}(Y(1), Y(0), S(1), S(0)|D, G = O)$ and $\tilde{P}(Y(1), Y(0), S(1), S(0)|Z, D, G = E)$ to complete the argument. Focus first on $\tilde{P}(Y(1), Y(0), S(1), S(0)|D, G = O)$ and its marginal $\tilde{P}(Y(d), S(d)|D, G = O)$. For $d \in \{0, 1\}$, let:

$$\begin{aligned} &\tilde{P}(Y(d) \in B_Y, S(d) \in B_S|D = d, G = O) = P(Y \in B_Y, S \in B_S|D = d, G = O) \\ &\tilde{P}(Y(d) \in B_Y|S(d) \in B_S, D \neq d, G = O) = \begin{cases} P(Y \in B_Y|S \in B_S, D = d, G = O), & \text{if } P(S \in B_S, D = d|G = O \\ 0, & \text{otherwise} \end{cases} \\ &\tilde{P}(S(d) \in B_S|D \neq d, G = O) = \frac{\gamma_d(B_S) - P(S \in B_S, D = d|G = O)}{P(D \neq d|G = O)}. \end{aligned} \tag{37}$$

This fully specifies $\tilde{P}(Y(d), S(d)|D, G = O)$. Note that $\tilde{P}(S(d)|G = O) = \gamma_d$, which is compatible with observational data by (36). Then, for all Borel sets $B_Y^d \in \mathcal{B}(\mathcal{Y})$ and $B_S^d \in \mathcal{B}(\mathcal{S})$ with $d \in \{0, 1\}$ let:

$$\begin{aligned} &\tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0|D, G = O) = \\ &\tilde{P}(Y(1) \in B_Y^1, S(1) \in B_S^1|D, G = O)\tilde{P}(Y(0) \in B_Y^0, S(0) \in B_S^0|D, G = O). \end{aligned} \tag{38}$$

Now, it is immediate that:

$$P(Y \in B_Y, S \in B_S, D = d|G = O) = P(Y \in B_Y, S \in B_S|D = d, G = O)P(D = d|G = O)$$
$$= \tilde{P}(Y(d) \in B_Y, S(d) \in B_S|D = d, G = O)\tilde{P}(D = d|G = O). \tag{39}$$

so $\tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0, D = d|G = O)$ induces the observational data distribution $P(Y \in B_Y^d, S \in B_S^d, D = d|G = O)$.

Next, consider $\tilde{P}(Y(1), Y(0), S(1), S(0)|Z, D, G = E)$ and its marginal $\tilde{P}(Y(d), S(d)|Z, D, G = E)$. Let:

$$\tilde{P}(Y(d) \in B_Y|S(d) \in B_S, Z \in B_Z, G = E) = \begin{cases} \tilde{P}(Y(d) \in B_Y|S(d) \in B_S, G = O) \text{ if } \tilde{P}(S(d) \in B_S|G = O) > 0 \\ 0, \text{ otherwise} \end{cases}$$
$$\tilde{P}(S(d) \in B_S|Z \in B_Z, D \neq d, G = E) = \begin{cases} \frac{\gamma_d(B_S) - P(S \in B_S, D = d|Z \in B_Z, G = E)}{P(D \neq d|Z \in B_Z|, G = E)}, \text{ if } P(D \neq d, Z \in B_Z|G = E) > 0 \\ 0, \text{ otherwise} \end{cases} \tag{40}$$

Note that $\tilde{P}(S(d) \in B_S|Z, G = E) = \gamma_d(B_S)$ P-a.s., and hence $\tilde{P}(S(d) \in B_S|G = E) = \gamma_d(B_S)$ which is compatible with experimental data by (36). Finally, let:

$$\begin{aligned} &\tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0|D, Z, G = E) = \\ &\tilde{P}(Y(1) \in B_Y^1, S(1) \in B_S^1|D, Z, G = O)\tilde{P}(Y(0) \in B_Y^0, S(0) \in B_S^0|D, Z, G = O) \end{aligned} \tag{41}$$

Then it follows that:

$$P(S \in B_S, D = d, Z \in B_Z|G = E) = P(S \in B_S|D = d, Z \in B_Z, G = E)P(D = d, Z \in B_Z|G = E) \tag{42}$$
$$= \tilde{P}(S(d) \in B_S|Z \in B_Z, D = d, G = E)\tilde{P}(D = d, Z \in B_Z|G = E) \tag{43}$$

so $\tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0, D = d, Z \in B_Z|G = E)$ induces the experimental data distribution $P(Y \in B_Y^d, S \in B_S^d, D = d, Z \in B_Z|G = E)$. Note that the joint distribution also satisfies Assumption RA.

To complete the proof observe that:

$$\begin{aligned}\tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0 | G = E) = \\ \tilde{P}(Y(1) \in B_Y^1, Y(0) \in B_Y^0, S(1) \in B_S^1, S(0) \in B_S^0 | G = O).\end{aligned} \tag{44}$$

□

This establishes that there are no additional features of the observed data that can be used to reject Assumption EV. Whenever we do not reject the testable implications, it is feasible that the distributions of unobservables and the treatment selection mechanisms are such that they satisfy the assumption. In that sense, the testable implications are the strongest possible. See Kitagawa (2015) for a similar result in the LATE setting.

It is instructive to consider a set of testable implications to which Proposition 4 does not apply. Consider:

$$\begin{aligned}E[S(1)|G = O] = E[S(1)|G = E] \\ E[S(0)|G = O] = E[S(0)|G = E].\end{aligned} \tag{45}$$

Again, $E[S(d)|G = g]$ may be bounded or point-identified following similar reasoning we used for $P(S(d)|G = g)$. We can then test whether there exists a value such that it is consistent with the identified sets for both $E[S(d)|G = O]$ and $E[S(d)|G = E]$. If not, we would reject (45) which would lead us to conclude that Assumption EV fails. If we do not reject it, then again, we cannot conclude that Assumption EV holds. However, unlike for the previous testable implication, there exist observed data distributions that satisfy (45) but there are no underlying distributions of potential outcomes and selection mechanisms that would satisfy Assumption EV.

For example, let there be perfect compliance $Z = D|G = E$ P-a.s. and suppose that $P(D = 0|G = O) = 1$ so that everyone in the observational data is untreated. Then $P[S(d)|G = E] = P[S|D = d, G = E]$ and $P[S(0)|G = O] = P[S|D = 0, G = O]$. Here $P(S(d)|G = g)$ for $d \in \{0, 1\}$ and $P(S(0)|G = g)$ are point-identified for any $g \in \{O, E\}$, since there is no selection. A fortiori, so are $E[S(d)|G = E]$ and $E[S(0)|G = O]$. Given that nobody in the observational data gets $D = 1$, $P(S(1)|G = O)$ and hence $E[S(1)|G = O]$, are unindentified. Suppose that $E[S|D = 0, G = E] = E[S|D = 0, G = O]$ which, along with the fact that $E[S(1)|G = O]$ is not identified, would mean that (45) holds. Thus, the testable implications regarding the mean do not refute Assumption EV. But the observed data may still be such that $P(S \in B|D = 0, G = E) \neq P(S \in B|D = 0, G = O)$ for some closed set B. Then the sharp implication (35) would still fail, refuting Assumption EV.

This illustrates the intuition of Proposition 4. Hence, we focus on testing (35).

TBA

A.4 Discretization of Short-term Outcomes

In this section, I clarify the implications of discretizing short-term outcomes. To this end, let a researcher pose a surjective discretization function $\lambda : \mathcal{S} \to \mathcal{S}^D := \{1, 2, \ldots, k\}$ for some $k < \infty$, and define $S^D(d) = \lambda(S(d))$. Note that this subsumes the case in which $S(d)$ is finitely supported, since then $\lambda(s) = s$ for all $s \in \mathcal{S}$. I introduce λ to clarify the subtle differences in applications of results of Section 5 when $S(d)$ is finitely supported and discretized. Similarly define discretized temporal link functions $m_d^D : \mathcal{S}^D \to \mathcal{Y}$, given by $m_d^D = E[Y(d)|S^D(d)] = E[Y(d)|\lambda(S(d))]$, and let $m^D = (m_0^D, m_1^D)$. Pose the following analog of Assumption MA under the discretization.

Assumption MA:D. *Suppose $\mathcal{M}^A$ and $\mathcal{M}^D$ are known or identified sets, and that $m \in \mathcal{M}^A \subseteq \mathcal{M}$. Then λ is such that $m^D \in \mathcal{M}^D$.*

Assumptions MA and MA:D are closely related. The former maintains that the researcher imposes some modeling assumption that will restrict feasible m, as in Section 2.3. The latter strengthens this notion and assumes that additionally m^D satisfies known restrictions after discretization. Of course, if Assumption MA holds for a finitely supported $S(d)$, then Assumption MA:D trivially follows by taking λ to be an identity function. The remark below explains that for some modeling assumptions and discretization functions, MA:D follows immediately from MA, but that it may be restrictive for others.

Remark 11. Consider LIV which states that $E[Y(d)|S(d) = s]$ is in $\mathcal{M}^A$ which contains only non-decreasing temporal link functions. Then $E[Y(d)|S^D = s]$ must also be non-decreasing for any order-preserving λ, so Assumption MA:D holds for an appropriately chosen λ. However, LUC states that $m_d(s) = E_O[Y|S = s, D = d]$, which does not directly imply that $m_d^D(s) = E[Y|S^D = s, D = d]$. A similar remark can be made for treatment invariance.

If $S(d)$ is finitely supported, MA and MA:D are equivalent and Section 5 characterizes the sharp identified set. If $S(d)$ is discretized and Assumption MA:D holds as a direct consequence of Assumption MA, such as under LIV, then results characterize the identified set $\mathcal{H}(\tau)$ that is sharp *under finitely-supported short-term outcomes*.[17] This is also the case if the researcher believes the modeling assumption holds under discretized data, i.e., is willing to maintain MA:D directly. Otherwise, the results in Section 5 should be viewed as providing an approximation of the sharp identified set.

17. Note that this set may be larger than the intractable identified set that would have been obtained using non-discretized data.

A.5 Results with Explicit Covariates

TBA

A.6 ATT and ATU

TBA

Preliminaries and Notation

I denote random variables and vectors using capital letters (e.g. Y), their cumulative distribution functions (CDF) by F (e.g. F_Y), laws by P (e.g. P_Y). I denote random sets with boldface letters (e.g. $\mathbf{Y}$), their capacity functionals by T (e.g. $T_{\mathbf{Y}}$) and containment functionals by C (e.g. $C_{\mathbf{Y}}$). I use $\overset{d}{=}$ to denote that a random element has a law, or an equivalent distribution-determining functional. (e.g. $Y \overset{d}{=} P_Y$ and $\mathbf{Y} \overset{d}{=} C_{\mathbf{Y}}$) Supports of random variables are denoted by corresponding script letters (e.g. $\mathcal{Y}$) and the support is defined as the smallest closed set containing the random variable $P-$a.s. Whenever there is no contextual ambiguity, I also use capital letters to denote sets. (usually A, B and K) Denote by $\mathcal{K}(A)$, $\mathcal{C}(A)$, $\mathcal{O}(A)$, $\mathcal{B}(A)$ the families of all compact, closed, open and Borel subsets of the set A, respectively. Let $co(A)$ be the closed convex hull of the set A. The sharp identified sets for a generic parameter θ is denoted by $\mathcal{H}(\theta)$, and for the distribution function (i.e. the law or equivalently the CDF) of a generic random vector Y by $\Gamma(Y)$ and $\mathcal{H}(P_Y)$. Let the set of distribution functions of random variables with support $\mathcal{Y}$ be $\mathcal{P}^{\mathcal{Y}}$. We assume throughout that $\mathcal{Y} \times \mathcal{S}$ is a locally compact, second countable Hausdorff space, more precisely $\mathbb{R}^{1+d}$ endowed with its natural topology, while any of its subspaces inherit their relative topologies.

In the proofs for simpler notation we will use the following random variable:

$$\tilde{Z} = \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1). \tag{46}$$

Observe that Assumptions RA and EV imply $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))|G$ and $(Y(d), S(d)) \perp\!\!\!\perp G$. Thus $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$.

Appendix B Supporting Results

B.1 Random Set Preliminaries

I briefly introduce the necessary concepts, and refer the reader to Molchanov (2017) and Molchanov and Molinari (2018) for a textbook treatment of the topic. More concise overviews are available in Beresteanu, Molchanov, and Molinari (2012) and Molchanov and Molinari (2014).

Define $\mathbf{X} : \Omega \to \mathcal{C}(\mathbb{R}^d)$ to be a measurable correspondence recalling that $\mathcal{C}(\mathbb{R}^d)$ is the collection of all closed subsets of $\mathbb{R}^d$.[18] We refer to $\mathbf{X}$ as a *random (closed) set*. Define the *containment functional* $C_{\mathbf{X}} : \mathcal{C}(\mathbb{R}^d) \to [0,1]$ of $\mathbf{X}$ as $C_{\mathbf{X}}(B) = P(\mathbf{X} \subseteq B)$, and the *capacity functional* $T_{\mathbf{X}} : \mathcal{K}(\mathbb{R}^d) \to [0,1]$ of $\mathbf{X}$ as $T_{\mathbf{X}}(K) = P(X \cap K \neq \emptyset)$, recalling that $\mathcal{K}(\mathbb{R}^d)$ is the collection of all compact subsets of $\mathbb{R}^d$. A *selection* of a random set $\mathbf{X}$ is a random vector X defined on the same probability space such that $P(X \in \mathbf{X}) = 1$. The set of all selections of $\mathbf{X}$ is denoted by $Sel(\mathbf{X})$. The set of all random vectors $X \in Sel(\mathbf{X})$ such that $E[||X||] < \infty$ is denoted by $Sel^1(\mathbf{X})$. Artstein's inequalities (Artstein (1983, Theorem 2.1), Beresteanu, Molchanov, and Molinari (2012, Theorem 2.1)) give an equivalent characterization of the set of distributions of all selections of a random set.

Lemma 2. *(Artstein's Inequalities) A probability distribution μ on a locally compact second countable Hausdorff space $\mathfrak{X}$ is the distribution of a selection of a random closed set $\mathbf{X}$ on the same space if and only if:*

$$\forall B \in \mathfrak{F}_{cont} : \ \mu(B) \geq C_{\mathbf{X}}(B) \Leftrightarrow \forall K \in \mathfrak{F}_{cap} : \ \mu(K) \leq T_{\mathbf{X}}(K) \tag{47}$$

where $\mathfrak{F}_{cont} \in \{\mathcal{C}(\mathfrak{X}), \mathcal{O}(\mathfrak{X})\}$ and $\mathfrak{F}_{cap} \in \{\mathcal{C}(\mathfrak{X}), \mathcal{O}(\mathfrak{X}), \mathcal{K}(\mathfrak{X})\}$. If $\mathbf{X}$ is almost surely compact, then (47) is equivalent to:

$$\forall K \in \mathcal{K}(\mathfrak{X}) : \ \mu(K) \geq C_{\mathbf{X}}(K). \tag{48}$$

Proof. For proof see Molchanov and Molinari (2018, Theorem 2.13, Corollary 2.14). □

If (47) holds for a distribution function P_X, then we call P_X *selectionable* with respect to the distribution of $\mathbf{X}$. μ is selectionable if and only if there exists a random element $X' \overset{d}{=} P_X$ and a random set $\mathbf{X}' \overset{d}{=} \mathbf{X}$ defined on the same probability space such that $P(X' \in \mathbf{X}') = 1$. Family of all distributions that satisfy (47) are called the *core* of the capacity $T_{\mathbf{X}}$. A family of compact sets $\mathcal{K}_{CD} \subseteq \mathcal{K}(\mathfrak{X})$ is a *core-determining class* if $\forall K \in \mathcal{K}_{CD} : \ \mu(K) \leq T_{\mathbf{X}}(K)$ implies (47). A core-determining class may reduce the number of conditions that need to be verified to consider μ selectionable.

If $\mathbf{X}$ has at least one integrable selection, that is $Sel^1(\mathbf{X}) \neq \emptyset$, then $\mathbf{X}$ is an *integrable random set*. Whenever the random variable $||\mathbf{X}|| = sup\{||X|| : X \in Sel(\mathbf{X})\}$ is integrable $E[||\mathbf{X}||] < \infty$, then $\mathbf{X}$ is said to be *integrably bounded*.

Definition 4. (Aumann Expectation) The Aumann expectation of an integrable random set $\mathbf{X}$

18. $\mathbf{X}$ is measurable if for every compact set $K \in \mathcal{K}(\mathbb{R}^d)$: $\{\omega \in \Omega : \mathbf{X}(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$. The codomain of the map $\mathbf{X}$ is equipped by the $\sigma-$algebra generated by the families of sets $\{B \in \mathcal{C}(\mathbb{R}^d) : B \cap K \neq \emptyset\}$ over $K \in \mathcal{K}(\mathbb{R}^d)$.

is defined as:

$$\mathbb{E}(\mathbf{X}) = cl\{E[X] : X \in Sel^1(\mathbf{X})\}. \tag{49}$$

If $\mathbf{X}$ is integrably bounded, then:

$$\mathbb{E}(\mathbf{X}) = \{E[X] : X \in Sel(\mathbf{X})\}. \tag{50}$$

Note that when $\mathbf{X}$ is a finite-dimensional and integrably bounded, $\mathbb{E}(\mathbf{X})$ is a closed set, and the closure operator is not used in the definition. (Molchanov (2017, Theorem 2.1.37))

The *support function* for a convex set $A \in \mathbb{R}^{d_A}$ is defined as $h_A(u) = \sup_{a \in A} a'u$ for $u \in \mathbb{R}^{d_A}$. The convex set A is uniquely determined by its its support function via intersections of all half-spaces defined by h_A as:

$$A = \bigcap_{u \in \mathbb{R}^{d_A} : ||u|| = 1} \{a \in \mathbb{R}^{d_A} : a'u \leq h_A(u)\}. \tag{51}$$

If $\mathbf{X}$ is integrably bounded and if either the underlying probability space is non-atomic, or if $\mathbf{X}$ is almost surely convex, then $h_{\mathbb{E}(\mathbf{X})}(u) = E[h_{\mathbf{X}}(u)]$ for all $u \in \mathbb{R}^{d_{\mathbf{X}}}$. (Molchanov and Molinari (2018, Theorem 3.11))

Recalling that $(\Omega, \mathcal{F}, P)$ is the underlying probabilty space, let $\mathcal{F}_0 \subsetneq \mathcal{F}$ be some sub-σ-algebra.

Definition 5. (Conditional Aumann Expectation) Let $\mathbf{X}$ be an integrable random set. For each sub-σ-algebra $\mathcal{F}_0 \subsetneq \mathcal{F}$, the conditional Aumann expectation of $\mathbf{X}$ given $\mathcal{F}_0$ is the $\mathcal{F}_0$-measurable random set $\mathbb{E}[\mathbf{X}|\mathcal{F}_0]$ such that:

$$Sel^1(\mathbb{E}[\mathbf{X}|\mathcal{F}_0], \mathcal{F}_0) = cl\{E[X|\mathcal{F}_0] : X \in Sel^1(\mathbf{X})\} \tag{52}$$

where $Sel^1(\cdot, \mathcal{F}_0)$ denote the set of integrable selections measurable with respect to $\mathcal{F}_0$ and the closure is taken in L^1.

For any integrable random set $\mathbf{X}$, the conditional Aumann expectation $\mathbb{E}[\mathbf{X}|\mathcal{F}_0]$ is integrable, unique and exists. If $\mathbf{X}$ is integrably bounded, so is $\mathbb{E}[\mathbf{X}|\mathcal{F}_0]$ (Molchanov (2017, Theorem 2.1.71)). When $\mathcal{F}_0$ is countably generated, then $cl\{E[X|\mathcal{F}_0] : X \in Sel^1(\mathbf{X})\} = \{E[X|\mathcal{F}_0] : X \in Sel^1(\mathbf{X})\}$. (Molchanov (2017, pp. 271), Li and Ogura (1998, Theorem 1)) Recall that when $\mathcal{F}_0$ is a σ-algebra generated by a random vector, it is countably generated. Therefore, for any random vector W, $Sel^1(\mathbb{E}[\mathbf{X}|\sigma(W)], \sigma(W))$ is a closed set.

If for all $A \in \mathcal{F}$ with $P(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)$ with positive probability, then the probability measure is said to have not atoms over

$\mathcal{F}_0$. Then, $\mathbb{E}[\mathbf{X}|\mathcal{F}_0]$ is almost surely convex and $\mathbb{E}[\mathbf{X}|\mathcal{F}_0] = \mathbb{E}[co(\mathbf{X})|\mathcal{F}_0]$ a.s. (Molchanov (2017, Theorem 2.1.77)) Then, $h_{\mathbb{E}[\mathbf{X}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{X})|\mathcal{F}_0]}(u) = E[h_{co(\mathbf{X})}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{X}}$. (Molchanov (2017, Theorem 2.1.72))[19] Note that this will hold for any sub-σ-algebra $\mathcal{F}_0$ by Lemma 3 under Assumption PS.

B.2 Other Known Results for Reference

Theorem 5. *Let E, F be metrizable and let G be any topological vector space. If E is a Baire space or if E is barreled and G is locally convex, then every separately equicontinuous family B of bilinear mappings of $E \times F$ into G is equicontinuous.*

Proof. See Schaefer and Wolff (1999, Theorem III.5.1). □

Corollary 3. *Let E, F be metrizable and let G be any topological vector space. If E is a Baire space or if E is barreled and G is locally convex, then every separately continuous bilinear map of $E \times F$ into G is continuous. (see also Treves (2016, pp. 425))*

Proof. Direct from Theorem 5. □

Appendix C Proofs

C.1 Auxiliary Lemmas

Lemma 3. *Suppose the probability space $(\Omega, \mathcal{F}, P)$ is non-atomic and that $\mathcal{F}_0 \subseteq \mathcal{F}$ is a sub-σ-algebra. Then P is atomless over $(\Omega, \mathcal{F}_0)$. That is, for all $A \in \mathcal{F}$ with $P(A) > 0$ there exists $B \in \mathcal{F}$ with $B \subseteq A$ such that $0 < P(B|\mathcal{F}_0) < P(A|\mathcal{F}_0)$ with positive probability.*

Proof. Pick any $A \in \mathcal{F}$ with positive measure and fix any $B \in \mathcal{F}$ such that $B \subseteq A$ and $0 < P(B) < P(A)$. B exists since $(\Omega, \mathcal{F}, P)$ is non-atomic. Let $C = A \setminus B$ and observe that $A = B \cup C$ and $B \cap C = \emptyset$. Thus, $P(A) = P(B) + P(C)$, $P(C) > 0$ and $P(A|\mathcal{F}_0) = P(B|\mathcal{F}_0) + P(C|\mathcal{F}_0)$ a.s. We proceed by way of contradiction supposing that $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. or $P(B|\mathcal{F}_0) = 0$ a.s. Consider $P(B|\mathcal{F}_0) = P(A|\mathcal{F}_0)$ a.s. first. Then, $P(C|\mathcal{F}_0) = 0$ which implies $P(C) = 0$, contradicting $P(C) > 0$. If $P(B|\mathcal{F}_0) = 0$ a.s., then $P(B) = 0$ a.s., contradicting $P(B) > 0$. Therefore, the set $\{\omega \in \Omega : 0 < P(B|\mathcal{F}_0)(\omega) < P(A|\mathcal{F}_0)(\omega)\}$ has positive probability, which concludes the proof. □

19. Theorem 2.1.72 states that $h_{\mathbb{E}[\mathbf{X}|\mathcal{F}_0]}(u) = E[h_\mathbf{X}(u)|\mathcal{F}_0]$ a.s. for all $u \in \mathbb{R}^{d_\mathbf{X}}$. If one wishes to use the support function to determine elements of $\mathbb{E}[\mathbf{X}|\mathcal{F}_0]$, the step $h_{\mathbb{E}[\mathbf{X}|\mathcal{F}_0]}(u) = h_{\mathbb{E}[co(\mathbf{X})|\mathcal{F}_0]}(u)$ by Theorem 2.1.77 is necessary.

Lemma 4. *Suppose that Assumption EV holds, and that experimental data are unobserved. Then the sharp identified set for the distribution function $P(Y(d), S(d))$ is:*

$$\Gamma^O(Y(d), S(d)) = \{\delta \in \mathcal{P}^{\mathcal{Y}\times\mathcal{S}} : \delta(B) \geq P_O((Y,S) \in B, D = d) \ \forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S})\} \tag{53}$$

Proof. Define the random set for $d \in \{0,1\}$:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y}\times\mathcal{S}, \text{ otherwise} \end{cases} . \tag{54}$$

$(\mathbf{Y}(d), \mathbf{S}(d))$ summarizes all observable information on $(Y(d), S(d))$ by definition. We wish to characterize the set of all selections $(Y(d), S(d)) \in Sel(\mathbf{Y}(d), \mathbf{S}(d))$ which expresses all available information contained in the observational data and assumptions. By Lemma 2, we have that a distribution function characterizes a selection in $Sel(\mathbf{Y}(d), \mathbf{S}(d))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{Y}\times\mathcal{S}): \quad P((Y(d), S(d)) \in B) \geq P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B). \tag{55}$$

For $B = \mathcal{Y}\times\mathcal{S}$, $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq K) = 1$.[20] For any other closed subset B, the containment functional can be written as:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B) &= P((Y(d), S(d)) \in B, D = d, G = O) = P((Y(d), S(d)) \in B, D = d | G = O) \\ &= P((Y,S) \in B, D = d | G = O). \end{aligned}$$

where the second equality follows by the fact that experimental data is unobserved so $P(G = O) = 1$, and the third by definition of Y and S. Hence, the identified set for $P(Y(d), S(d))$ follows by (47). Sharpness follows by construction since the random set expresses all available information from the data and assumptions. □

Lemma 5. *Suppose that Assumptions RA and EV hold. Then the sharp identified set* $\Gamma(Y(d), S(d))$

20. The support of a random variable X is the smallest closed set $\mathcal{X}$ such that $P(X \in \mathcal{X}) = 1$. Hence $\mathcal{Y}\times\mathcal{S} \in \mathcal{C}(\mathcal{Y}\times\mathcal{S})$.

for the distribution function $P(Y(d), S(d))$ *is:*

$$\Gamma(Y(d), S(d)) =$$

$$\left\{ \delta \in \mathcal{P} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P(S \in B_S, D = d | Z, G = E), P(S \in B_S, D = d | G = O)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P((Y,S) \in B, D = d | G = O) \end{array} \right] \end{array} \right\} \tag{56}$$

Proof. Define the random set for $d \in \{0, 1\}$:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y,S)\}, \text{ if } (D,G) = (d,O) \\ \mathcal{Y} \times \{S\}, \text{ if } (D,G) = (d,E) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases} . \tag{57}$$

Let $\tilde{I}$ be the set of triples random elements (E_1, E_2, E_3) such that $(E_1, E_2, E_3) \in \mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}$ and $(E_1, E_2) \perp\!\!\!\perp E_3$. We wish to characterize the set of selections $(Y(d), S(d), \tilde{Z}) \in Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$ which expresses all available information from the data and assumptions. When Assumptions RA and EV hold, $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I} \neq \emptyset$.

By Lemma 2, we have that a distribution function characterizes a selection in $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S} \times \tilde{\mathcal{Z}}) : \quad P((Y(d), S(d), \tilde{Z}) \in B) \geq P((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}) \subseteq B) \tag{58}$$

By Molchanov and Molinari (2018, Theorem 2.33), (58) is equivalent to:

$$\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \quad P((Y(d), S(d)) \in B | \tilde{Z}) \geq P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B | \tilde{Z}) \ P\text{-a.s.} \tag{59}$$

Possible forms that B can take are: 1) $B = \mathcal{Y} \times \mathcal{S}$; 2) $B = \mathcal{Y} \times B_S$ for some $B_S \subsetneq \mathcal{S}$; 3) $B \neq \mathcal{Y} \times B_S$ for all $B_S \subseteq \mathcal{S}$. Now consider the containment functional $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B | \tilde{Z})$ for each case.

For $B = \mathcal{Y} \times \mathcal{S}$, $P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B | \tilde{Z}) = 1$ P-a.s. If $B = \mathcal{Y} \times B_S$ for some $B_S \subsetneq \mathcal{S}$, then

P-a.s.:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) &= P((Y,S) \in B, D = d|\tilde{Z}) \\ &= P(Y \in \mathcal{Y}, S \in B_S, D = d|\tilde{Z}) \\ &= P(S \in B_S, D = d|\tilde{Z}) \end{aligned}$$

where the first equality is by definition of the random set, the second is by the fact that $B = \mathcal{Y} \times B_S$, and the third by definition of $\mathcal{Y}$. Finally, if for all $B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S$:

$$P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = \begin{cases} 0, \text{ if } \tilde{Z} \in \mathcal{Z} \text{ (i.e. } G = E) \\ P((Y,S) \in B, D = d|\tilde{Z}), \text{ if } \tilde{Z} \notin \mathcal{Z} \text{ (i.e. } G = O) \end{cases}.$$

To see why the first case holds, define the section of B at point s as $B_Y(s) = \{y : (y,s) \in B\}$. Then observe that if for all $B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S$, we must have for some s that $B_Y(s) \subsetneq \mathcal{Y}$. Therefore whenever $G = E$ (or equivalently $\tilde{Z} \in \mathcal{Z}$), the random set $(\mathbf{Y}(d), \mathbf{S}(d)) = \mathcal{Y} \times \{S\} \not\subseteq B$. Hence, only if $G = O$ can the containment functional be positive, that is, when $\tilde{Z} \neq \mathcal{Z}$.

Thus we can collect the relevant cases to define the containment functional:

$$\begin{aligned} P((\mathbf{Y}(d), \mathbf{S}(d)) \subseteq B|\tilde{Z}) = &\begin{cases} P(S \in B_S, D = d|\tilde{Z}), \text{ if } B = \mathcal{Y} \times B_S \text{ for some } B_S \subsetneq \mathcal{S} \\ \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P((Y,S) \in B, D = d|G = O), \text{ otherwise} \end{cases} \\ = &\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ &\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P((Y,S) \in B, D = d|G = O) \end{aligned}$$

Hence, a distribution function characterizes a selection in $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z}))$ if and only if $\forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})$ P-a.s.:

$$P((Y(d), S(d)) \in B|\tilde{Z}) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] P(S \in B_S, D = d|\tilde{Z}) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] \mathbb{1}[\tilde{Z} \notin \mathcal{Z}] P((Y,S) \in B, D = d|G = O) \end{array} \right].$$

Finally, to incorporate the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d), S(d))$, we intersect $Sel((\mathbf{Y}(d), \mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$

to obtain:

$$P((Y(d),S(d))\in B) \geq ess\sup_{\tilde{z}} \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] P(S\in B_S, D=d|\tilde{Z})+ \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S]\mathbb{1}[\tilde{Z}\notin \mathcal{Z}] P((Y,S)\in B, D=d|G=O) \end{array} \right].$$

$$= \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S] ess\sup_{\tilde{Z}} P(S\in B_S, D=d|\tilde{Z})+ \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P((Y,S)\in B, D=d|G=O) \end{array} \right]$$

$$= \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]\times \\ \max\left(ess\sup_{Z} P(S\in B_S, D=d|Z,G=E), P(S\in B_S, D=d|G=O)\right)+ \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S] P((Y,S)\in B, D=d|G=O) \end{array} \right]$$

where the first line follows by the fact that $\tilde{Z} \perp\!\!\!\perp (Y(d),S(d))$, the second by the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y}\times B_S]$ and $\mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y}\times B_S]$ refer to mutually exclusive events, and the third by definition of $\tilde{Z}$ and the fact that $P(G=g)>0$ for $g\in\{O,E\}$.

□

Lemma 6. *Let* $\Gamma^O(Y(d))$ *and* $\Gamma(Y(d))$ *be the sets of marginals of distributions in* $\Gamma^O(Y(d),S(d))$ *and* $\Gamma(Y(d),S(d))$*. Then:*

$$\Gamma^O(Y(d)) = \Gamma(Y(d)) = \{\delta\in\mathcal{P}^{\mathcal{Y}} : \delta(B)\geq P(Y\in B, D=d|G=O)\ \forall B\in\mathcal{C}(\mathcal{Y})\}. \tag{60}$$

Proof. For any Borel set $B\in\mathcal{B}(\mathbb{R})$, by definition of a marginal distribution function we have:

$$P(Y(d)\in B) = P(Y(d)\in B, S(d)\in\mathcal{S}) = P((Y(d),S(d))\in B\times\mathcal{S}) \tag{61}$$

where the last line is by equivalence of events $\{Y(d)\in B, S(d)\in\mathcal{S}\}$ and $\{(Y(d),S(d))\in B\times\mathcal{S}\}$. Lemma 4 yields the sharp identified set for joint distributions $P(Y(d),S(d))$ using only observational data:

$$\Gamma^O(Y(d),S(d)) = \{\delta\in\mathcal{P} : \delta(B)\geq P((Y,S)\in B, D=d|G=O)\ \forall B\in\mathcal{C}(\mathcal{Y}\times\mathcal{S})\}. \tag{62}$$

Hence the sharp identified set for marginals $P(Y(d))$ using only observational data is:

$$\begin{aligned}\Gamma^O(Y(d)) &= \{P(Y(d))\in\mathcal{P}^{\mathcal{Y}} : \exists\delta\in\Gamma^O(Y(d),S(d)) \text{ s.t } P(Y(d)\in B)=\delta(B\times\mathcal{S})\ \forall B\in\mathcal{B}(\mathbb{R})\} \\ &= \{P(Y(d))\in\mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B)\geq P((Y,S)\in B\times\mathcal{S}, D=d|G=O)\ \forall B\in\mathcal{C}(\mathcal{Y})\} \\ &= \{P(Y(d))\in\mathcal{P}^{\mathcal{Y}} : P(Y(d)\in B)\geq P(Y\in B, D=d|G=O)\ \forall B\in\mathcal{C}(\mathcal{Y})\}\end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 4 and the third is by (61).

Lemma 5 yield the sharp identified set for joint distributions $P(Y(d), S(d))$ using combined data:

$$\Gamma(Y(d), S(d)) =$$
$$\left\{ \delta \in \mathcal{P} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P(S \in B_S, D = d | Z, G = E), P(S \in B_S, D = d | G = O)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P((Y, S) \in B, D = d | G = O) \end{array} \right] \end{array} \right\} \tag{63}$$

Observe that when we define the marginals, we only need to look at Borel sets of the form $B \times \mathcal{S}$ with $B \subsetneq \mathcal{Y}$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 0$ in the expression above. Thus, the sharp identified set for marginals $P(Y(d))$ using combined data is:

$$\begin{aligned} \Gamma(Y(d)) &= \{P(Y(d)) \in \mathcal{P}^{\mathcal{Y}} : \exists \delta \in \Gamma(Y(d), S(d)) \text{ s.t } P(Y(d) \in B) = \delta(B \times \mathcal{S}) \ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{P(Y(d)) \in \mathcal{P}^{\mathcal{Y}} : P(Y(d) \in B) \geq P((Y, S) \in B \times \mathcal{S}, D = d | G = O) \ \forall B \in \mathcal{C}(\mathcal{Y})\} \\ &= \{P(Y(d)) \in \mathcal{P}^{\mathcal{Y}} : P(Y(d) \in B) \geq P(Y \in B, D = d | G = O) \ \forall B \in \mathcal{C}(\mathcal{Y})\} \end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 5 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 0$, and the third is by (61). □

Remark 12. The formulation of the sharp identified sets $\Gamma^O(Y(d))$ and $\Gamma(Y(d))$ coincides by application of (47) to the random set:

$$\mathbf{Y}(d) = \begin{cases} \{Y\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y}, \text{ otherwise} \end{cases}.$$

Lemma 7. *Let $\Gamma^O(S(d))$ and $\Gamma(S(d))$ be the sets of marginals of distributions in $\Gamma^O(Y(d), S(d))$ and $\Gamma(Y(d), S(d))$. Then:*

$$\Gamma^O(S(d)) = \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P(S \in B, D = d | G = O) \ \forall B \in \mathcal{C}(\mathcal{S})\} \tag{64}$$

$$\Gamma(S(d)) = \left\{ \delta \in \mathcal{P}^{\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{S}) : \\ \delta(B) \geq \max\left(ess\sup_Z P(S \in B, D = d | Z, G = E), P(S \in B, D = d | G = O)\right) \end{array} \right\} \tag{65}$$

Let $\Gamma^E(S(d))$ be the sharp identified set for $P(S(d)|G = E)$ obtained using only experimental

data. Then:

$$\Gamma^E(S(d)) = \{\delta \in \mathcal{P}^{\mathcal{S}} : ess\sup_Z P(S \in B, D = d|Z, G = E)\ \forall B \in \mathcal{C}(\mathcal{S})\}. \tag{66}$$

Proof. For any Borel set $B \in \mathcal{B}(\mathbb{R})$, by definition of a marginal distribution function we have:

$$P(S(d) \in B) = P(S(d) \in \mathcal{Y}, S(d) \in B) = P((Y(d), S(d)) \in \mathcal{Y} \times B) \tag{67}$$

where the last line is by equivalence of events $\{Y(d) \in \mathcal{Y}, S(d) \in B\}$ and $\{(Y(d), S(d)) \in \mathcal{Y} \times B\}$. Lemma 4 yields the sharp identified set for joint distributions $P(Y(d), S(d))$ using only observational data:

$$\Gamma^O(Y(d), S(d)) = \{\delta \in \mathcal{P} : \delta(B) \geq P((Y, S) \in B, D = d|G = O)\ \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S})\}. \tag{68}$$

Hence the sharp identified set for marginals $P(S(d))$ using only observational data is:

$$\begin{aligned}
\Gamma^O(S(d)) &= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \Gamma^O(Y(d), S(d)) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P((Y, S) \in \mathcal{Y} \times B, D = d|G = O)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\
&= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : P(S(d) \in B) \geq P(S \in B, D = d|G = O)\ \forall B \in \mathcal{C}(\mathcal{S})\}
\end{aligned}$$

where the first line is by definition of a marginal distribution, second is by Lemma 4 and the third is by (67).

Lemma 5 yield the sharp identified set for joint distributions $P(Y(d), S(d))$ using combined data:

$$\Gamma(Y(d), S(d)) =$$
$$\left\{ \delta \in \mathcal{P} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{Y} \times \mathcal{S}) : \\ \delta(B) \geq \left[\begin{array}{l} \mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] \times \\ \max\left(ess\sup_Z P(S \in B_S, D = d|Z, G = E), P(S \in B_S, D = d|G = O)\right) + \\ \mathbb{1}[\forall B_S \subseteq \mathcal{S} : B \neq \mathcal{Y} \times B_S] P((Y, S) \in B, D = d|G = O) \end{array} \right] \end{array} \right\} \tag{69}$$

Observe that when we define the marginals, we only need to look at Borel sets of the form $\mathcal{Y} \times B$, which means that for all sets of interest $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$ in the expression above.

Thus, the sharp identified set for marginals $P(S(d))$ using combined data is:

$$\begin{aligned}\Gamma(S(d)) &= \{P(S(d)) \in \mathcal{P}^{\mathcal{S}} : \exists \delta \in \Gamma(Y(d), S(d)) \text{ s.t } P(S(d) \in B) = \delta(\mathcal{Y} \times B)\ \forall B \in \mathcal{B}(\mathbb{R})\} \\ &= \{\delta \in \mathcal{P}^{\mathcal{S}} : \delta(B) \geq P(S \in B, D = d | G = O)\ \forall B \in \mathcal{C}(\mathcal{S})\} \\ &= \left\{ \delta \in \mathcal{P}^{\mathcal{S}} : \begin{array}{l} \forall B \in \mathcal{C}(\mathcal{S}) : \\ \delta(B) \geq \max\left(ess\sup_Z P(S \in B_S, D = d | Z, G = E), P(S \in B_S, D = d | G = O)\right) \end{array} \right\}.\end{aligned}$$

where the first line is by definition of a marginal distribution, and the second is by Lemma 5 and the fact that $\mathbb{1}[\exists B_S \subseteq \mathcal{S} : B = \mathcal{Y} \times B_S] = 1$.

For $\Gamma^E(S(d))$ a simplified version of the argument for Lemma 5 applies, and is therefore omitted. □

Remark 13. The formulation of the sharp identified sets $\Gamma^O(S(d))$ and $\Gamma(S(d))$ coincides by application of (47) to the random set:

$$(\mathbf{S}(d), \tilde{Z}) = \begin{cases} \{S\}, \text{ if } (D, G) \in \{(d, E), (d, O)\} \\ \mathcal{S}, \text{ otherwise} \end{cases}$$

and finding the set of selections $Sel(\mathbf{S}(d), \tilde{Z}) \cap I_S$ where I_S is the set of random elements (E_1, E_2) such that $E_1 \perp\!\!\!\perp E_2$.

Lemma 8. *Let $\mathcal{Y}$ be a compact set. If there exists $d \in \{0, 1\}$ such that $V_O[Y|S, D = d] > 0$ $P-$a.s., then $E_O[Y|S, D = d] \in (\inf \mathcal{Y}, \sup \mathcal{Y})$ $P-$a.s.*

Proof. I prove that $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. and $E_O[Y|S, D = d] > \inf \mathcal{Y}$ $P-$a.s follows by a symmetric argument. Since $\mathcal{Y}$ is a compact set, both $\sup \mathcal{Y}$ and $\inf \mathcal{Y}$ are finite.

By contraposition suppose that $P(E_O[Y|S, D = d] \geq \sup \mathcal{Y}) > 0$. Then by definition of $\mathcal{Y}$, $P(E_O[Y|S, D = d] = \sup \mathcal{Y}) > 0$, so there exists a Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D = d) > 0$ such that $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Now we show that this implies

$P(Y = \sup \mathcal{Y} | S \in B, D = d) = 1$. Suppose not, so that $P(Y = \sup \mathcal{Y} | S \in B, D = d) < 1$, then:

$$\begin{aligned} E_O[Y|S \in B, D = d] =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d] P(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y \neq \sup \mathcal{Y}, S \in B, D = d] P_O(Y \neq \sup \mathcal{Y}|S \in B, D = d) \\ =& E_O[Y|Y = \sup \mathcal{Y}, S \in B, D = d] P_O(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ =& \sup \mathcal{Y} P_O(Y = \sup \mathcal{Y}|S \in B, D = d) + \\ & E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] P_O(Y < \sup \mathcal{Y}|S \in B, D = d) \\ <& \mathcal{Y} \end{aligned} \tag{70}$$

where the first equality is by LIE, second is by definition of $\mathcal{Y}$, third by $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$, and the fourth by $E_O[Y|Y < \sup \mathcal{Y}, S \in B, D = d] < \sup \mathcal{Y}$ and $P(Y = \mathcal{Y}|S \in B, D = d) < 1$. By assumption we had that $E_O[Y|S \in B, D = d] = \sup \mathcal{Y}$. Then (70) yields a contradiction, showing that $P(Y = \sup \mathcal{Y}|S \in B, D = d) = 1$. But then $V_O[Y|S \in B, D = d] = 0$ and $P_O(S \in B|D = d) > 0$, so $P(V_O[Y|S \in B, D = d] = 0) > 0$ which contradicts $V_O[Y|S \in B, D = d] > 0$ $P-$a.s. Thus $V_O[Y|S, D = d] > 0$ $P-$a.s. implies $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. □

Lemma 9. *Let ς_d be an arbitrary selection of $\mathbf{S}(d)$ with the distribution γ_d. Then for any $B \in \mathcal{B}(\mathbb{R})$ such that $\gamma_d(B) > 0$:*

$$P_O(D = d|\varsigma_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)}. \tag{71}$$

Proof. Fix an arbitrary $B \in \mathcal{B}(\mathbb{R})$ such that $\gamma_d(B) > 0$. Then:

$$\begin{aligned} P_O(D = d|\varsigma_d \in B) &= \frac{P_O(\varsigma_d \in B|D = d) P_O(D = d)}{P(\varsigma_d \in B)} \\ &= \frac{P_O(\varsigma_d \in B|D = d) P_O(D = d)}{\gamma_d(B)} \\ &= \frac{P_O(S \in B|D = d) P_O(D = d)}{\gamma_d(B)} \\ &= \frac{P_O(S \in B, D = d)}{\gamma_d(B)} \end{aligned}$$

where the first line is by Bayes' theorem, second is by $\varsigma_d \overset{d}{=} \gamma_d$, third is by the fact that $\varsigma_d \in Sel(\mathbf{S}(d))$ and $\mathbf{S}(d) = \{S\}$ when $D = d$, and the last is by observation. □

Lemma 10. *Suppose Assumption RA holds. Assume that there is perfect experimental compliance so* $Z = D|G = E$ $P-a.s.$ *and define conditions:*

C.1 (Surrogacy) $Y \perp\!\!\!\perp D|S, G = E$;

C.2 (Comparability) $Y \perp\!\!\!\perp G|S$.

Then:

i) C.1 implies $E_E[Y(1)|S(1) = s] = E_E[Y(0)|S(0) = s]$ *for all* $s \in \mathcal{S}$;

ii) C.1 and EV imply $E_g[Y(1)|S(1) = s] = E_{g'}[Y(0)|S(0) = s]$ *for all* $s \in \mathcal{S}$ *and* $g, g' \in \{O, E\}$;

iii) C.2 implies $E_O[Y|S = s] = E_E[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_E[Y(0)|S(0) = s]P_E(D = 0|S = s)$ *for all* $s \in \mathcal{S}$;

iv) C.2 and EV imply $E_O[Y|S = s] = E_g[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_{g'}[Y(0)|S(0) = s]P_E(D = 0|S = s)$ *for all* $s \in \mathcal{S}$ *and* $g, g' \in \{O, E\}$;

v) C.1 and C.2 imply $E_O[Y|S = s] = E_E[Y(d)|S(d) = s]$ *for all* $s \in \mathcal{S}$;

vi) C.1, C.2 and EV imply $E_O[Y|S = s] = E_g[Y(d)|S(d) = s]$ *for all* $s \in \mathcal{S}$ *and* $g \in \{O, E\}$.

Proof. i) Write for any $d \in \{0, 1\}$:

$$E_E[Y|S] = E_E[Y|S, D = d] = E_E[Y(d)|S(d), D = d] = E_E[Y(d)|S(d)] \tag{72}$$

where the first equality is by surrogacy, second is by definition, and third is by random assignment and perfect compliance.

ii) Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *i)*.

iii) Write:

$$\begin{aligned} E_O[Y|S = s] &= E_E[Y|S = s] \\ &= E_E[Y(1)|S(1) = s, D = 1]P_E(D = 1|S = s) \\ &+ E_E[Y(0)|S(0) = s, D = 0]P_E(D = 0|S = s) \\ &= E_E[Y(1)|S(1) = s]P_E(D = 1|S = s) + E_E[Y(0)|S(0) = s]P_E(D = 0|S = s) \end{aligned} \tag{73}$$

where the first equality is by comparability, second is by LIE and definitions of Y and S, and the third is by random assignment and perfect compliance. *iv)* Under Assumption EV, $E_E[Y(d)|S(d)] = E[Y(d)|S(d)]$. The result then follows from *iii)*.

v) Immediate from *i)* and *iii)*.

vi) Immediate from *v)* under Assumption EV.

□

C.2 Main Results

Proposition 1. *Suppose Assumptions RA and EV hold. Then:*

i) $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$;

ii) $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$.

Proof of Proposition 1. I show that $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)}))$, which immediately yields $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$.

Let $\Pi(\nu_0, \nu_1)$ be the set of couplings of probability measures ν_0 and ν_1 defined as (Villani et al. (2009, Definition 1.1)):

$$\Pi(\nu_0, \nu_1) = \left\{ \delta \in \mathcal{P}^{\mathcal{Y}} \times \mathcal{P}^{\mathcal{Y}} : \ \forall A \subseteq \mathcal{Y} \quad \begin{array}{l} \delta(A \times \mathcal{Y}) = \nu_0(A), \\ \delta(\mathcal{Y} \times A) = \nu_1(A) \end{array} \right\}. \tag{74}$$

$\Pi(\nu_0, \nu_1)$ is always non-empty. (Galichon (2018, Section 2.1)) The data never reveal $(Y(0), Y(1))$ jointly, so the sharp identified set for $P_{Y(0),Y(1)}$ given $P(Y(1))$ and $P(Y(0))$ is simply the set of couplings of pushworward measures $P(Y(1))$ and $P(Y(0))$. Using the sharp identified sets for the marginals $P(Y(d)) \in \Gamma^O(Y(d))$ for $d \in \{0, 1\}$, the sharp identified set $\mathcal{H}^O(P_{Y(0),Y(1)})$ is then the union of all possible couplings:

$$\mathcal{H}^O(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1) \in \Gamma^O(Y(0)) \times \Gamma^O(Y(1))} \Pi(\nu_0, \nu_1). \tag{75}$$

Similarly for $\mathcal{H}(P_{Y(0),Y(1)})$:

$$\mathcal{H}(P_{Y(0),Y(1)}) = \bigcup_{(\nu_0,\nu_1) \in \Gamma(Y(0)) \times \Gamma(Y(1))} \Pi(\nu_0, \nu_1). \tag{76}$$

Lemma 6 shows that $\Gamma^O(Y(d)) = \Gamma(Y(d))$ for any $d \in \{0, 1\}$. That $\mathcal{H}^O(P_{Y(0),Y(1)}) = \mathcal{H}(P_{Y(0),Y(1)})$ follows.

Next, observe that τ is a functional of $P_{Y(0),Y(1)}$. It is then immediate that $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$, □

Lemma 1. *(Nested Misspecification) Let $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$ be misspecified identified sets for some parameter τ. Let d be the point-to-set distance defined as $d(A, t) := \inf \{\|t - a\| : a \in A\}$ for*

$A \subseteq \mathbb{R}$ and $t \in \mathbb{R}$. Then:

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) \leq d(\tilde{\mathcal{H}}, \tau)$$

Proof of Lemma 1.

$$d(\tilde{\mathcal{H}}^{O/A}, \tau) = \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}^{O/A}\right\} \leq \inf\left\{\|t-\tau\| : t \in \tilde{\mathcal{H}}\right\} = d(\tilde{\mathcal{H}}, \tau) \tag{77}$$

where the inequality follows by $\tilde{\mathcal{H}} \subseteq \tilde{\mathcal{H}}^{O/A}$. □

Theorem 1. *Let Assumptions RA, EV, and MA hold, and $\tilde{Z} := \mathbb{1}[G = E]Z + \mathbb{1}[G = O](sup\mathcal{Z} + 1) \in \tilde{\mathcal{Z}}$. The sharp identified set for (m, γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z}) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)),\ \varsigma_d \overset{d}{=} \gamma_d,\ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d]\ a.s. \end{array} \right\}. \tag{14}$$

where I is the set of random elements $(E_1, E_2) \in \mathcal{S} \times \tilde{\mathcal{Z}}$ such that $E_1 \perp\!\!\!\perp E_2$.

Proof of Theorem 1. I first find the set of all $(S(0), S(1), Y(0), Y(1))$ which are consistent with the data, Assumptions RA and EV and $E[|Y(d)|] < \infty$. This follows by adapting and extending arguments in the proofs of Beresteanu, Molchanov, and Molinari (2012, Propositions 2.4, 2.5). By definition, this then yields the set of all corresponding (m, γ). $\mathcal{H}(m, \gamma)$ follows by collecting pairs of (m, γ) such that $m \in \mathcal{M}^A$.

Step 1: Restrictions on selections without modeling assumptions

Recall the definition of the random set:

$$(\mathbf{Y}(d), \mathbf{S}(d)) = \begin{cases} \{(Y, S)\}, \text{ if } (D, G) = (d, O) \\ \mathcal{Y} \times \{S\}, \text{ if } (D, G) = (d, E) \\ \mathcal{Y} \times \mathcal{S}, \text{ otherwise} \end{cases} \tag{78}$$

which summarizes all information on $(Y(d), S(d))$ contained in the data, by definition. Observe that $(\mathbf{Y}(d), \mathbf{S}(d)) = \mathbf{Y}(d) \times \mathbf{S}(d)$. Events $\{\omega \in \Omega : \mathbf{S}(0) \subsetneq \mathcal{S}\}$ and $\{\omega \in \Omega : \mathbf{S}(1) \subsetneq \mathcal{S}\}$ are mutually exclusive, as are $\{\omega \in \Omega : \mathbf{Y}(0) \subsetneq \mathcal{Y}\}$ and $\{\omega \in \Omega : \mathbf{Y}(1) \subsetneq \mathcal{Y}\}$. All information on $(S(0), S(1), Y(0), Y(1))$ contained in the data can thus be summarized by $(S(0), S(1), Y(0), Y(1)) \in Sel(\mathbf{S}(0) \times \mathbf{S}(1) \times \mathbf{Y}(0) \times \mathbf{Y}(1))$. By observation, this is equivalent to:

$$(S(0), S(1), Y(0), Y(1), \tilde{Z}) \in Sel(\mathbf{S}(0)) \times Sel(\mathbf{S}(1)) \times Sel(\mathbf{Y}(0)) \times Sel(\mathbf{Y}(1)) \tag{79}$$

By definition of $\tilde{Z}$ and, Assumptions RA and EV, $\tilde{Z} \perp\!\!\!\perp (S(0), S(1), Y(0), Y(1))$. Moreover, by assumption $E[|Y(d)|] < \infty$. Let $\tilde{I}$ be the set of all random elements $(E_1, E_2, E_3, E_4, E_5) \in$

$\mathcal{S}^2 \times \mathcal{Y}^2 \times \tilde{\mathcal{Z}}$ such that $E_5 \perp\!\!\!\perp (E_1, E_2, E_3, E_4)$ and $E[|E_j|] < \infty$ for $j \in \{3, 4\}$. All information in the data, Assumptions RA and EV, and $E[|Y(d)|] < \infty$ can be expressed by:

$$\begin{aligned}(S(0), S(1), Y(0), Y(1), \tilde{Z}) &\in Sel(\mathbf{S}(0)) \times Sel(\mathbf{S}(1)) \times Sel^1(\mathbf{Y}(0)) \times Sel^1(\mathbf{Y}(1)) \times \{\tilde{Z}\} \cap \tilde{I} \\ &:= \mathcal{H}^{EV/RA}((S(0), S(1), Y(0), Y(1), \tilde{Z}))\end{aligned} \tag{80}$$

where the second line follows by observation.

Step 2: Restrictions on (m, γ) without the modeling assumption

By definition, the set of all (m, γ) consistent with the data, Assumptions RA and EV and $E[|Y(d)|] < \infty$ is:

$$\begin{aligned}&\mathcal{H}^{EV/RA}(m, \gamma) \\ &= \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{M} \times (\mathcal{P}^{\mathcal{S}})^2 : \exists (\varsigma_0, \varsigma_1, \upsilon_0, \upsilon_1, \tilde{Z}) \in \mathcal{H}^{EV/RA}((S(0), S(1), Y(0), Y(1), \tilde{Z})), \\ \forall d \in \{0, 1\}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}.\end{aligned} \tag{81}$$

As an intermediate step, I show that this is equivalent to:

$$\begin{aligned}&\mathcal{H}'^{EV/RA}(m, \gamma) \\ &= \left\{ \begin{array}{l} (m, \gamma) \in \mathcal{M} \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0, 1\}, \\ \exists (\varsigma_d, \upsilon_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times Sel^1(\mathbf{Y}(d)) \times \{\tilde{Z}\} \cap \bar{I}, \;\; \gamma_d \overset{d}{=} \varsigma_d, \;\; m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d] \text{ a.s.} \end{array} \right\}.\end{aligned} \tag{82}$$

where $\bar{I}$ is the set of all random elements $(E_1, E_2, E_3) \in \mathcal{S} \times \mathcal{Y} \times \tilde{\mathcal{Z}}$ such that $E_3 \perp\!\!\!\perp (E_1, E_2)$.

Fix any $(m, \gamma) \in \mathcal{H}^{EV/RA}(m, \gamma)$. By definition there exists $(\varsigma_0, \varsigma_1, \upsilon_0, \upsilon_1, \tilde{Z})$ such that for each $d \in \{0, 1\}$, $\varsigma_d \in Sel(\mathbf{S}(d))$, $\upsilon_d \in Sel^1(\mathbf{Y}(d))$, $(\varsigma_d, \upsilon_d) \perp\!\!\!\perp \tilde{Z}$. Moreover, and $m_d(\varsigma_d) = E[\upsilon_d | \varsigma_d]$ a.s. and $\gamma_d \overset{d}{=} \varsigma_d$. Hence, $(m, \gamma) \in \mathcal{H}'^{EV/RA}(m, \gamma)$.

Next, fix any $(m, \gamma) \in \mathcal{H}'^{EV/RA}(m, \gamma)$. Take the corresponding $(\varsigma_d, \upsilon_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times Sel^1(\mathbf{Y}(d)) \times \{\tilde{Z}\}$ for $d \in \{0, 1\}$ that generate (m, γ). By Villani et al. (2009, Section 1)), one can always construct an independent coupling, so there exists $(\varsigma_0', \varsigma_1', \upsilon_0', \upsilon_1')$ such that $(\varsigma_d, \upsilon_d) \overset{d}{=} (\varsigma_d', \upsilon_d')$ for all d. Repeating this, one can create an independent coupling $(\varsigma_0'', \varsigma_1'', \upsilon_0'', \upsilon_1'', \tilde{Z}'')$ such that $(\varsigma_0', \varsigma_1', \upsilon_0', \upsilon_1') \overset{d}{=} (\varsigma_0'', \varsigma_1'', \upsilon_0'', \upsilon_1'')$, $\tilde{Z} \overset{d}{=} \tilde{Z}''$ and $(\varsigma_0'', \varsigma_1'', \upsilon_0'', \upsilon_1'') \perp\!\!\!\perp \tilde{Z}''$. But since $(\varsigma_d, \upsilon_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times Sel^1(\mathbf{Y}(d)) \times \{\tilde{Z}\}$ for $d \in \{0, 1\}$, by definition, there will exist an element in $\mathcal{H}^{EV/RA}((S(0), S(1), Y(0), Y(1), \tilde{Z}))$ equal in distribution to $(\varsigma_0'', \varsigma_1'', \upsilon_0'', \upsilon_1'', \tilde{Z}'')$. Hence, for any $d \in \{0, 1\}$ it will yield $\varsigma_d'' \overset{d}{=} \varsigma_d \overset{d}{=} \gamma_d$ and identical m_d up to almost sure equivalence. ???

...

$$\mathcal{H}^{EV/RA}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z}) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)),\ \varsigma_d \overset{d}{=} \gamma_d,\ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}. \tag{83}$$

Step 3: Identified set $\mathcal{H}(m,\gamma)$

It only remains to impose Assumption MA. To do so, observe that a valid identified set is:

$$\mathcal{H}(m,\gamma) = \mathcal{H}^{EV/RA}(m,\gamma) \cap (\mathcal{M} \times (\mathcal{P}^{\mathcal{S}})^2). \tag{84}$$

That it is sharp is immediate, since for every $(m,\gamma) \in \mathcal{H}(m,\gamma)$ there exist selections ς_d and υ_d that are consistent with the data and remaining assumptions. Therefore:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists \varsigma_d \in Sel((\mathbf{S}(d), \tilde{Z}) \cap I, \\ \exists \upsilon_d \in Sel^1(\mathbf{Y}(d)),\ \varsigma_d \overset{d}{=} \gamma_d,\ m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d] \text{ a.s.} \end{array} \right\}. \tag{85}$$

□

Theorem 2. *Let Assumptions RA, EV, MA, and PS hold. If* $\mathbf{Y}(d)$ *is integrably bounded, the sharp identified set for* (m,γ) *is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \varsigma_d \overset{d}{=} \gamma_d, \\ \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \\ \forall u \in \{-1,1\} : \ \ um_d(\varsigma_d) \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \ \ a.s. \end{array} \right\} \tag{16}$$

where $h_{co(\mathcal{Y})}(u) = \sup_{y \in co(\mathcal{Y})} uy$ *and* $P_O(D = d|\varsigma_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)}$ *for all* $B \in \mathcal{B}(\mathcal{S})$ *with* $\gamma_d(B) > 0$.

Proof of Theorem 2. I first show that the restrictions on m_d given a selection ς_d can be equivalently restated using the conditional Aumann expectation. The further equivalent characterizations then follow by Artstein's theorem and convexification properties of conditional Aumann expectation on non-atomic probability spaces. I then equivalently characterize these restrictions such that they become invariant to the chosen selections ς_d up do their marginal distributions γ_d.

Step 1: Representation of restriction on γ_d

I follow similar steps to those in the proof of Lemma 5 to equivalently characterize the

condition $\gamma_d \overset{d}{=} \varsigma_d$. By Lemma 2, a distribution function characterizes a selection in $Sel((\mathbf{S}(d), \tilde{Z}))$ if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S} \times \tilde{\mathcal{Z}}): \quad P((S(d), \tilde{Z}) \in B) \geq P((\mathbf{S}(d), \tilde{Z}) \subseteq B) \tag{86}$$

$$\Leftrightarrow \forall B \in \mathcal{C}(\mathcal{S}): \quad P(S(d) \in B|\tilde{Z}) \geq P(\mathbf{S}(d) \subseteq B|\tilde{Z}) \text{ a.s.} \tag{87}$$

where the second line follows by Molchanov and Molinari (2018, Theorem 2.33). Now consider the containment functional $P(\mathbf{S}(d) \subseteq B|\tilde{Z})$. If $B = \mathcal{S}$, $P(\mathbf{S}(d) \subseteq B|\tilde{Z}) = 1$. If $B \subseteq \mathcal{S}$, then $P(\mathbf{S}(d) \subseteq B|\tilde{Z}) = P(S \subseteq B, D = d|\tilde{Z})$. Hence, $\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z}))$ such that $\gamma_d \in \mathcal{P}^{\mathcal{S}}$ and $\gamma_d \overset{d}{=} \varsigma_d$ if and only if $\forall B \in \mathcal{C}(\mathcal{S})$, $B \neq \mathcal{S}$:

$$P(\varsigma_d \in B|\tilde{Z}) \geq P(S \subseteq B, D = d|\tilde{Z}) \text{ a.s.}$$

Finally, to incorporate restrictions imposed by assumptions, intersect $Sel(\mathbf{S}(d), \tilde{Z}) \cap I$. Thus, $\exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I$ such that $\gamma_d \overset{d}{=} \varsigma_d$ if and only if $\forall B \in \mathcal{C}(\mathcal{S})$, $B \neq \mathcal{S}$:

$$\begin{aligned} P(\varsigma_d \in B) &\geq ess\sup_{\tilde{Z}} P(S \subseteq B, D = d|\tilde{Z}) \\ &= \max\left(ess\sup_{Z} P_E(S \in B_S, D = d|Z), P_O(S \in B_S, D = d)\right) \end{aligned} \tag{88}$$

where the first line follows by definition of I, and the second by definition of $\tilde{Z}$ and $P(G = g) > 0$ for $g \in \{O, E\}$ since we observe two datasets.

Step 2: Reformulating restrictions on m_d

Fix an arbitrary $d \in \{0, 1\}$ and ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel(\mathbf{S}(d)) \times \{\tilde{Z}\} \cap I$. Let $\sigma(\varsigma_d)$ be the sub-σ-algebra generated by ς_d. Let $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d] = cl\{E[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$, where the closure is taken in L^1 space of all $\sigma(\varsigma_d)$-measurable functions. $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ exists, is a unique random set, and has at least one integrable selection. Since $\mathbf{Y}(d)$ is integrably bounded, so is $\mathbb{E}[\mathbf{Y}(d)|S(d)]$. (Molchanov (2017, Theorem 2.1.71)) ς_d is a measurable selection, hence a random variable. Therefore, the conditioning sub-σ-algebra $\sigma(\varsigma_d)$ of $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ is generated by a random variable and is thus countably generated. Then since $\mathbf{Y}(d)$ is integrably bounded and defined on $\mathbb{R}$, $\{E[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$ is a closed set, so $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d] = \{E[\upsilon_d|\varsigma_d] : \upsilon_d \in Sel^1(\mathbf{Y}(d))\}$. (Li and Ogura (1998, Theorem 1)) It is then immediate that:

$$\exists \upsilon_d \in Sel^1(\mathbf{Y}(d)): \quad m_d(\varsigma_d) = E[\upsilon_d|\varsigma_d] \text{ a.s.} \Leftrightarrow m_d(\varsigma_d) \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d] \text{ a.s..} \tag{89}$$

Therefore:

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} m \in \mathcal{M}^A, \gamma \in (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d,\ m_d \in Sel^1(\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\} \tag{90}$$

$$= \left\{ \begin{array}{l} m \in \mathcal{M}^A, \gamma \in (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\},\ \exists(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap I, \\ \gamma_d \overset{d}{=} \varsigma_d,\ m_d \in Sel(\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]) \end{array} \right\} \tag{91}$$

where the second line follows since $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ is integrably bounded.

By Assumption PS and Lemma 3, P has no atoms over ς_d for any ς_d. Since $E[|Y(d)|] < \infty$ for all $d \in \{0,1\}$, $\mathbf{Y}(d)$ is integrable. Thus, $\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ is almost surely convex and equal to $\mathbb{E}[co(\mathbf{Y}(d))|\varsigma_d]$. (Molchanov (2017, Theorem 2.1.77)) Therefore, $h_{\mathbb{E}[\mathbf{Y}|\varsigma_d]}(u) = h_{\mathbb{E}[co(\mathbf{Y}(d))|\varsigma_d]}(u)$ a.s. for all $u \in \mathbb{R}$. By $\mathbb{E}[co(\mathbf{Y}(d))|\varsigma_d] = \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ and integrability of the latter, the former set is also integrable. It follows that $h_{\mathbb{E}[co(\mathbf{Y}(d))|\varsigma_d]}(u) = E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$ (Molchanov (2017, Theorem 2.1.72)). Hence, $h_{\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]}(u) = E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$.

We then have:

$$\begin{aligned} & m_d \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d] \text{ a.s.} \\ \Leftrightarrow & um_d(\varsigma_d) \leq h_{\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]}(u) \text{ a.s. } \forall u \in \{-1,1\} \\ \Leftrightarrow & um_d \leq E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] \text{ a.s. } \forall u \in \{-1,1\} \end{aligned} \tag{92}$$

where the second line is by Rockafellar (1970, Theorem 13.1), and the third is by $h_{\mathbb{E}[\mathbf{Y}(d)|\varsigma_d]}(u) = E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d]$ a.s. for all $u \in \mathbb{R}$. Moreover:

$$\begin{aligned} E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d] &= E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D=d]P_O(D=d|\varsigma_d) + E[h_{co(\mathbf{Y}(d))}(u)|\varsigma_d, D\neq d]P(D\neq d|\varsigma_d) \\ &= uE[Y|\varsigma_d, D=d]P_O(D=d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D\neq d|\varsigma_d) \\ &= uE[Y|S, D=d]P_O(D=d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D\neq d|\varsigma_d) \end{aligned} \tag{93}$$

where the first line is by LIE, the second is by definition of $\mathbf{Y}(d)$, and the final is by observing that since $\varsigma_d \in Sel(\mathbf{S}(d))$, it must be $P(\varsigma_d = S|D=d) = 1$ by definition of $\mathbf{S}(d)$. Finally, by Lemma 9, we have for any $B \in \mathcal{B}(\mathcal{S})$ such that $\gamma_d(B) > 0$, $P(D=d|\varsigma_d \in B) = \frac{P_O(S\in B, D=d)}{\gamma_d(B)}$.

Step 3: Equivalent representation of $\mathcal{H}(m,\gamma)$

Define the set:

$$\mathcal{H}^I = \left\{ \begin{array}{l} m \in \mathcal{M}^A, \gamma \in (\mathcal{P}^{\mathcal{S}})^2 : \forall d \in \{0,1\}, \\ \forall B \in \mathcal{C}(\mathcal{S}) : \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S \in B, D = d|Z), P_O(S \in B, D = d)\right), \varsigma_d \overset{d}{=} \gamma_d, \\ \forall u \in \{-1,1\} : \ \ um_d \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \text{ a.s.} \end{array} \right\} \tag{94}$$

where $P(D = d|\varsigma_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)}$ for any $B \in \mathcal{B}(\mathcal{S})$ such that $\gamma_d(B) > 0$.

Now we show that $\mathcal{H}(m, \gamma) = \mathcal{H}^I$. First, pick $(m, \gamma) \in \mathcal{H}(m, \gamma)$. Then $m \in \mathcal{M}^A$, and $\forall d \in \{0,1\}$ there exists ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$, $\gamma_d \overset{d}{=} \varsigma_d$ and $m_d \in \mathbb{E}[\mathbf{Y}(d)|\varsigma_d]$ a.s.. Then by (88), (92), (93), and Lemma 9 $(m, \gamma) \in \mathcal{H}^I$.

Conversely, pick $(m, \gamma) \in \mathcal{H}^I$. Then $m \in \mathcal{M}^A$, and we know by (88) and Lemma 2 that for every $d \in \{0,1\}$ there exists ς_d such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$, $\gamma_d \overset{d}{=} \varsigma_d$. It only remains to show that $m_d \in Sel^1(\mathbb{E}[\mathbf{Y}(d)|\varsigma_d])$.[21] We observe that for any $\varsigma'_d \overset{d}{=} \varsigma_d \overset{d}{=} \gamma_d$, $P(D = d|\varsigma'_d) = P(D = d|\varsigma_d)$ a.s. This is immediate since for any Borel set $B \in \mathcal{B}(\mathcal{S})$ such that $\gamma_d(B) > 0$: $P(D = d|\varsigma'_d \in B) = \frac{P_O(S \in B, D = d)}{\gamma_d(B)} = P(D = d|\varsigma_d \in B)$. We can then write:

$$\forall u \in \{-1,1\} : \ \ um_d \leq uE[Y|S, D = d]P_O(D = d|\varsigma'_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma'_d) \text{ a.s.} \tag{95}$$

$$\Leftrightarrow \forall u \in \{-1,1\} : \ \ um_d \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \text{ a.s.} \tag{96}$$

where the second line follows by $P(D = d|\varsigma'_d) = P(D = d|\varsigma_d)$ a.s. Since m_d satisfies (95) for some random variable $\varsigma'_d \overset{d}{=} \gamma_d$ by assumption, by retracing (93) and (92), we have that $m_d \in Sel^1(\mathbb{E}[\mathbf{Y}(d)|\varsigma_d])$ for ς_d. Hence $\mathcal{H}^I = \mathcal{H}(m, \gamma)$.

□

Theorem 3. *Let Assumptions RA, EV, MA, PS and BS hold. Suppose $\mathcal{S}$ is a finite set and that $\mathcal{M}^A$ is closed and convex. Then:*

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma}) \in \mathcal{H}(m,\gamma)} T(\tilde{m}, \tilde{\gamma}) \right].$$

Proof of Theorem 3. I first show that $T(m, \gamma)$ is a jointly continuous functional. I then prove

21. Note that Lemma 2 does not imply that every random variable $\varsigma_d \overset{d}{=} \gamma_d$ is a selection such that $(\varsigma_d, \tilde{Z}) \in Sel((\mathbf{S}(d), \tilde{Z})) \cap \tilde{I}$, only that such a selection exists and has distribution γ_d.

that the set:

$$\mathcal{H}^{WC}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M} \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ m_d(s) \leq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + 1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}, \\ m_d(s) \geq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} \end{array} \right\}. \tag{97}$$

is compact and convex. Then, I demonstrate that $\mathcal{H}(m,\gamma) = \mathcal{H}^{WC}(m,\gamma) \cap \left(\mathcal{M}^A \times (\Delta(k))^2\right)$ which is compact, convex, and non-empty. Finally, since the continuous image in $\mathbb{R}$ over a non-empty compact-convex set is a closed interval, the identified set is a closed interval.

Step 1: T is jointly continuous.

Endow the set of reals with its natural topology, making it a locally convex topological vector space (t.v.s.). By bilinearity of the Riemann-Stieltjes integral in the integrand and integrator, $T(m,\gamma)$ is a bilinear map. Since T is a bilinear map in a finite-dimensional space, it is separately continuous in each argument. Note that $T : \mathbb{R}^{2d_s} \times \mathbb{R}^{2d_s} \to \mathbb{R}$ and that $\mathbb{R}^{2d_s}$ is Polish (separable and completely metrizable), and hence metrizable. By a corollary of the first Baire category theorem, every Polish space is a Baire space, so $\mathbb{R}^{2d_s}$ is a Baire space. (Willard (2004, Corollary 25.4)) By Corollary 3, T is jointly continuous since every separately continuous bilinear map from a product of a Baire space and a metrizable space to a locally convex t.v.s. is jointly continuous.

Step 2: $\mathcal{H}^{WC}(m,\gamma)$ is convex.

Pick any $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$ and fix $a \in (0,1)$. I show $a(m,\gamma) + (1-a)(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$. It is immediate that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$a\gamma_d + (1-a)\gamma'_d \geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right) \tag{98}$$

since both γ_d and γ'_d satisfy the same condition. Next, note that for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} &\frac{a\gamma'_d(s) + (1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)} - \frac{1}{a\gamma_d(s) + (1-a)\gamma'_d(s)} \\ &= \frac{(a\gamma'_d(s) + (1-a)\gamma_d(s))(a\gamma_d(s) + (1-a)\gamma'_d(s)) - \gamma_d(s)\gamma'_d(s)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\ &= \frac{(a^2 + (1-a)^2 - 1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2 + \gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\ &= \frac{2a(a-1)\gamma'_d(s)\gamma_d(s) + a(1-a)(\gamma'_d(s)^2 + \gamma_d(s)^2)}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \\ &= \frac{a(1-a)(\gamma_d(s) - \gamma_d(s)')^2}{\gamma_d(s)\gamma'_d(s)(a\gamma_d(s) + (1-a)\gamma'_d(s))} \geq 0. \end{aligned} \tag{99}$$

Then for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} am_d(s) + (1-a)m'_d(s) &\geq E_O[Y|S=s,D=d]P_O(S=s,D=d)\left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) \\ &\geq E_O[Y|S=s,D=d]P_O(S=s,D=d)\left(\frac{a\gamma'_d(s)+(1-a)\gamma_d(s)}{\gamma_d(s)\gamma'_d(s)}\right) \\ &\geq E_O[Y|S=s,D=d])\frac{P_O(S=s,D=d)}{a\gamma_d(s)+(1-a)\gamma'_d(s)} \end{aligned} \tag{100}$$

where the first line follows by $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$, second is by observation and the third is by (99). Finally, for any $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} am_d(s) + (1-a)m'_d(s) &\leq (E_O[Y|S=s,D=d]-1)\,P_O(S=s,D=d)\left(\frac{a}{\gamma_d(s)} + \frac{1-a}{\gamma'_d(s)}\right) + 1 \\ &\leq (E_O[Y|S=s,D=d]-1)\,\frac{P_O(S=s,D=d)}{a\gamma_d(s)+(1-a)\gamma'_d(s)} + 1 \end{aligned} \tag{101}$$

where $(m,\gamma),(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$ yields the first line, and the second follows by (99) and $(E_O[Y|S=s,D=d]-1) \leq 0$. Hence, $a(m,\gamma)+(1-a)(m',\gamma') \in \mathcal{H}^{WC}(m,\gamma)$.

Step 3: $\mathcal{H}^{WC}(m,\gamma)$ is compact.

It is immediate that $\mathcal{H}^{WC}(m,\gamma)$ is bounded since $\mathcal{H}^{WC}(m,\gamma) \subseteq [0,1]^k \times (\Delta(k))^2$, by Assumption BS and $\mathcal{S}$ being a finite set. I now show it is closed, which proves compactness since (m,γ) is finite-dimensional. Take any convergent sequence $(m^j,\gamma^j) \to (m^*,\gamma^*)$ as $j \to \infty$. Then for any j, $d \in \{0,1\}$ and $s \in \mathcal{S}$:

$$\begin{aligned} \gamma^j_d(s) &\geq \max\left(ess\sup_Z P_E(S=s,D=d|Z), P_O(S=s,D=d)\right) \\ m^j_d(s) &\leq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma^j_d(s)} + 1 - \frac{P_O(S=s,D=d)}{\gamma^j_d(s)} \\ m^j_d(s) &\geq E_O[Y|S=s,D=d]\frac{P_O(S=s,D=d)}{\gamma^j_d(s)}. \end{aligned} \tag{102}$$

Then by continuity of $f(x) = 1/x$ and limit-preservation of weak-inequalities:

$$\begin{aligned}
\gamma_d^*(s) &\geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s, D=d)\right) \\
m_d^*(s) &\leq E_O[Y|S=s, D=d]\frac{P_O(S=s, D=d)}{\gamma_d^*(s)} + 1 - \frac{P_O(S=s, D=d)}{\gamma_d^*(s)} \\
m_d^*(s) &\geq E_O[Y|S=s, D=d]\frac{P_O(S=s, D=d)}{\gamma_d^*(s)}.
\end{aligned} \tag{103}$$

Hence $(m^*, \gamma^*) \in \mathcal{H}^{WC}(m, \gamma)$, so the set is closed. Then it is also compact.

Step 4: $\mathcal{H}(m, \gamma)$ is compact, convex, and non-empty.

By assumption, $\mathcal{M}^A$ is convex and compact. $\Delta(k)$ is a $k-$dimensional simplex, and hence convex and compact. Then $\mathcal{M}^A \times (\Delta(k))^2$ is also convex and compact. By Assumption BS, $h_{co(\mathcal{Y})}(-1) = 0$ and $h_{co(\mathcal{Y})}(1) = 1$. By Proposition 2, it is then immediate that $\mathcal{H}(m, \gamma) = \mathcal{H}^{WC}(m, \gamma) \cap \mathcal{M}^A \times (\Delta(k))^2$. Since $\mathcal{H}^{WC}(m, \gamma)$ is compact and convex, and intersections preserve compactness and convexity, $\mathcal{H}(m, \gamma)$ is compact and convex. If maintained assumptions hold, then $\mathcal{H}(m, \gamma)$ is non-empty.

Step 5: $\mathcal{H}(\tau)$ is an interval.

T was shown to be a continuous map, so it preserves connectedness. Hence, $\mathcal{H}(\tau) = \{T(m, \gamma) : (m, \gamma) \in \mathcal{H}(m, \gamma)\}$ is a connected set. Since $\mathcal{H}(\tau) \subseteq \mathbb{R}$, it is an interval. Continuous images preserve compactness, so the $\mathcal{H}(\tau)$ is a compact interval, so:

$$\mathcal{H}(\tau) = \left[\inf_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m, \gamma), \sup_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m, \gamma)\right] \tag{104}$$

$$= \left[\min_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m, \gamma), \max_{(m,\gamma)\in\mathcal{H}(m,\gamma)} T(m, \gamma)\right] \tag{105}$$

where the second line follows by continuity of T and compactness of $\mathcal{H}(m, \gamma)$.

□

Corollary 1. *Let conditions of Theorem 3 hold. If $\mathcal{H}(m|\cdot)$ has minimal and maximal selectors with respect to T, then:*

$$\mathcal{H}(\tau) = \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(L_{\tilde{\gamma}}, \tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} T(U_{\tilde{\gamma}}, \tilde{\gamma})\right].$$

Proof of Corollary 1. By observation, it is immediate that iterated and joint minima and maxima

search over all values of the set $\mathcal{H}(m,\gamma)$, so we can write:

$$\mathcal{H}(\tau) = \left[\min_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma}), \max_{(\tilde{m},\tilde{\gamma})\in\mathcal{H}(m,\gamma)} T(\tilde{m},\tilde{\gamma})\right] \tag{106}$$

$$= \left[\min_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \min_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma}), \max_{\tilde{\gamma}\in\mathcal{H}(\gamma)} \max_{\tilde{m}\in\mathcal{H}(m|\tilde{\gamma})} T(\tilde{m},\tilde{\gamma})\right]. \tag{107}$$

By definition of L_γ and U_γ:

$$\forall \tilde{m} \in \mathcal{H}(m|\gamma): \;\; T(L_\gamma,\gamma) \leq T(\tilde{m},\gamma) \leq T(U_\gamma,\gamma). \tag{108}$$

Therefore:

$$\begin{aligned} \min_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(L_\gamma,\gamma) \\ \max_{\tilde{m}\in\mathcal{H}(m|\gamma)} T(\tilde{m},\gamma) &= T(U_\gamma,\gamma). \end{aligned} \tag{109}$$

□

Proposition 2. *Let Assumptions RA, EV, MA, PS and BS hold. Suppose $\mathcal{S}$ is a finite set. Then the sharp identified set for (m,γ) is:*

$$\mathcal{H}(m,\gamma) = \left\{ \begin{array}{l} (m,\gamma) \in \mathcal{M}^A \times (\Delta(k))^2 : \forall d \in \{0,1\},\ \forall s \in \mathcal{S}, \\ \gamma_d(s) \geq \max\left(ess\sup_Z (P_E(S=s,D=d|Z), P_O(S=s,D=d)\right), \\ \forall u \in \{-1,1\}: \;\; u m_d(s) \leq u E[Y|S,D=d]\frac{P_O(S=s,D=d)}{\gamma_d(s)} + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S=s,D=d)}{\gamma_d(s)}\right) \end{array} \right\}. \tag{22}$$

Proof of Proposition 2. Since $\mathcal{S} = \{1,2,\ldots,k\}$, we can represent γ_d as an element of a $k-$dimensional simplex $\Delta(k)$ and $m_d \in \mathcal{Y}^k$. Let $\gamma_d(s)$ and $m_d(s)$ denote the $s-$th element of the corresponding vectors. By the proof of Theorem 2, γ_d is selectionable if and only if:

$$\forall B \in \mathcal{C}(\mathcal{S}): \gamma_d(B) \geq \max\left(ess\sup_Z P_E(S\in B, D=d|Z), P_O(S\in B, D=d)\right). \tag{110}$$

Since $\mathcal{S}$ is finite, by Beresteanu, Molchanov, and Molinari (2012, Lemma B.1) $\{\{s\}: s\in\mathcal{S}\}$ is a core-determining class. Then following the same steps as the proof of Theorem 2, we can show that γ_d is selectionable if and only if:

$$\forall s \in \mathcal{S}: \gamma_d(s) \geq \max\left(ess\sup_Z P_E(S=s, D=d|Z), P_O(S=s, D=d)\right). \tag{111}$$

Finally, fix any $\varsigma_d \overset{d}{=} \gamma_d$ such that for any Borel set $B \in \mathcal{B}(\mathcal{S})$ with $\gamma_d > 0$ we have $P_O(D = d|\varsigma_d \in B) = \frac{P_O(S\in B, D=d)}{\gamma_d(B)}$. Equivalently, we have $P_O(D = d|\varsigma_d = s) = \frac{P_O(S=s,D=d)}{\gamma_d(s)}$ $\forall s \in \mathcal{S}$. Then:

$$\begin{aligned}
&\forall u \in \{-1, 1\}: \quad um_d \leq uE[Y|S, D = d]P_O(D = d|\varsigma_d) + h_{co(\mathcal{Y})}(u)P(D \neq d|\varsigma_d) \text{a.s.} \\
\Leftrightarrow &\forall u \in \{-1, 1\}: \quad um_d \leq uE[Y|S, D = d]\frac{P_O(S = s, D = d)}{\gamma_d(s)} \\
&\qquad\qquad + h_{co(\mathcal{Y})}(u)\left(1 - \frac{P_O(S = s, D = d)}{\gamma_d(s)}\right) \quad \forall s \in \mathcal{S}.
\end{aligned} \tag{112}$$

□

Corollary 2. *Suppose Assumptions RA and EV hold. If $\mathcal{Y} = \mathbb{R}$, the sharp identified set for τ is $\mathcal{H}(\tau) = \mathbb{R}$. If we additionally maintain $\mathcal{Y} = [0, 1]$:*

$$\mathcal{H}(\tau) = [E_O[YD] - E_O[Y(1 - D)] - P_O(D = 0), E_O[YD] - E_O[Y(1 - D)] + P_O(D = 1)]. \tag{26}$$

In both cases, $0 \in \mathcal{H}(\tau)$ and the sign of τ not identified.

Proof of Corollary 2. Suppose first that Assumptions RA and EV hold. Assume that $\mathcal{Y} = \mathbb{R}$ and pick an arbitrary $\tilde{c} \in \mathbb{R}$. I show that $\tilde{c} \in \mathcal{H}^O(\tau)$ which is equivalent to $\tilde{c} \in \mathcal{H}(\tau)$ by Proposition 1.

Define a distribution function for any $(a, d) \in \mathbb{R} \times \{0, 1\}$:

$$\gamma_{a|d}(B) = P_O(Y \in B, D = d) + \mathbb{1}[a \in B]P_O(D \neq d)$$

for any Borel set $B \in \mathcal{B}(\mathcal{Y})$. Recall from Lemma 6 that:

$$\Gamma^O(Y(d)) = \{\gamma \in \mathcal{P}^{\mathcal{Y}} : \gamma(B) \geq P_O(Y \in B, D = d) \ \forall B \in \mathcal{C}(\mathcal{Y})\}.$$

Since $\mathcal{C}(\mathcal{Y}) \subseteq \mathcal{B}(\mathcal{Y})$, then $\gamma_{a|d}(B) \in \Gamma^O(Y(d))$ for any $(a, d) \in \mathbb{R} \times \{0, 1\}$. Note also that any coupling of $\gamma_{a|1} \in \Gamma^O(Y(1))$ and $\gamma_{a'|0} \in \Gamma^O(Y(0))$ is compatible with the observed data.

Next, observe that $\gamma_{a|d}$ is a pushforward measure of the random variable $Y\mathbb{1}[D = d] + a\mathbb{1}[D \neq d]$, which has the expectation of $E[Y\mathbb{1}[D = d]] + aP_O(D \neq d)$. Let $c = \frac{\tilde{c} - E_O[YD] + E_O[Y(1-D)]}{P_O(D\neq d)} \in \mathbb{R}$. Then $\gamma_{c|1}$ yields the expected value:

$$E[YD] + \frac{\tilde{c} - E_O[YD] + E_O[Y(1 - D)]}{P_O(D = 0)}P_O(D = 0) = \tilde{c} + E_O[Y(1 - D)].$$

Similarly, $\gamma_{0|0}$ yields the expected value $E[Y(1 - D)]$.

Now take $\gamma_{c|1} \in \Gamma^O(Y(1))$ and $\gamma_{0|0} \in \Gamma^O(Y(0))$ as distribution functions of $Y(1)$ and $Y(0)$, recalling that any coupling of $\gamma_{c|1}$ and $\gamma_{c|0}$ is compatible with the observed data. It follows that $\tau = E[Y(1) - Y(0)] = \tilde{c}$. Since $\tilde{c}$ was arbitrary, $\mathcal{H}^O(\tau) = \mathbb{R}$. By Proposition 1, $\mathcal{H}^O(\tau) = \mathcal{H}(\tau)$.

Next, let $\mathcal{Y} = [0, 1]$. Since Proposition 1 holds for any $\mathcal{Y} \subseteq \mathbb{R}$, we can again recover $\mathcal{H}(\tau)$ by using only distributions in $\Gamma^O(Y(d))$. Equivalently, we can find $\mathcal{H}(\tau)$ by utilizing only information in the observational data. Then, by elementary arguments as in Manski (1990), the bounds in (26) follow. □

Proposition 3. *Let Assumptions EV and LUC hold.*

i) Suppose the observed data distribution $P_O(Y, S, D)$ is such that $V_O[Y|S, D = d] > 0$ $P-$a.s. for some $d \in \{0, 1\}$ and that $\mathcal{Y}$ is a bounded set. Then $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

ii) If the observed data distribution $P_O(Y, S, D)$ is such that $E_O[Y|S, D = d]$ is a trivial measurable function for all $d \in \{0, 1\}$, then τ is point-identified, and $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$.

Proof of Proposition 3. We prove the claims in order.

i)

$\mathcal{Y}$ is closed by definition. Since it is bounded, it is a compact set. Then $\sup \mathcal{Y} < \infty$ and $\inf \mathcal{Y} > -\infty$. Using arguments of Manski (1990), the sharp upper bound of $\mathcal{H}^O(\tau)$ is:

$$\tau \leq E_O[Y(2D - 1)] + \sup \mathcal{Y} P_O(D = 1) - \inf \mathcal{Y} P_O(D = 0) = \sup \mathcal{H}^O(\tau). \tag{113}$$

By Lemma 8 $V_O[Y|S, D = d] > 0$ $P-$a.s. implies $E_O[Y|S, D = d] < \sup \mathcal{Y}$ $P-$a.s. If there exists $d \in \{0, 1\}$ s.t. $V[Y|S, D = d] > 0$ $P-$a.s., then it must be that for every Borel subset $B \subseteq \mathcal{B}(\mathcal{S})$ with $P_O(S \in B|D = d) > 0$ we have $E_O[Y|S \in B, D = d] < \sup \mathcal{Y}$. Under Assumption LUC we then have:

$$\begin{aligned} E_O[Y(d)|D \neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d]P_O(D \neq d) \\ &= E_O[E_O[Y(d)|S(d), D = d|D \neq d]]P_O(D \neq d) \\ &= E_O[E_O[Y|S, D = d]|D \neq d]P_O(D \neq d) \\ &< \sup \mathcal{Y} P_O(D \neq d) \end{aligned} \tag{114}$$

where the first line is by LIE, second by Assumption LUC, third by definition, and the fourth since $E[Y|S \in B, D = d] < \sup \mathcal{Y}$ for every Borel set B of positive measure. Then under LUC:

$$\begin{aligned} E[Y(d)] &= E_O[Y \mathbb{1}[D = d]] + E[Y(d)|D \neq d]P_O(D \neq d) \\ &< E_O[Y \mathbb{1}[D = d]] + \sup \mathcal{Y} P_O(D \neq d). \end{aligned} \tag{115}$$

Therefore, under Assumption LUC:

$$\begin{aligned}\tau &= E[Y(1) - Y(0)] \\ &= E_O[YD] + E[Y(1)|D=0]P_O(D=0) - E[Y(1-D)] - E[Y(0)|D=1]P_O(D=1) \\ &< E_O[Y(2D-1)] + \sup \mathcal{Y} P_O(D=1) - \inf \mathcal{Y} P_O(D=0) = \sup \mathcal{H}^O(\tau)\end{aligned}$$

where the inequality follows by (115). Thus $\sup \mathcal{H}^{O/LUC}(\tau) < \sup \mathcal{H}^O(\tau)$. So there must exist a point in $\mathcal{H}(\tau)$ which is not contained in $\mathcal{H}^{O/LUC}(\tau)$. We conclude that $\mathcal{H}^{O/LUC}(\tau) \subsetneq \mathcal{H}^O(\tau)$.

ii)

Suppose that for every $d \in \{0,1\}$ $E_O[Y|S, D=d]$ is a trivial measurable function. Hence there exists a $y \in \mathcal{Y}$ such that $E_O[Y|S,D] = y$ $P-$a.s.

Then, following the same steps as in (114):

$$\begin{aligned}E_O[Y(d)|\neq d] &= E_O[E_O[Y(d)|S(d), D \neq d]|D \neq d]P_O(D \neq d) \\ &= E_O[E_O[Y(d)|S(d), D = d|D \neq d]]P_O(D \neq d) \\ &= E_O[E_O[Y|S, D=d]|D \neq d]P_O(D \neq d) \\ &= yP_O(D \neq d)\end{aligned} \tag{116}$$

where the final line follows since $E_O[Y|S,D] = y$ $P-$a.s. and $Supp(\mathcal{S}(d)) = \mathcal{S}$. Given that y is identified by the data, then $E_O[Y(d)]$ is identified for every $d \in \{0,1\}$, so τ is too. It is also immediate that $\mathcal{H}(\tau) = \mathcal{H}^{O/LUC}(\tau)$ since for every $d \in \{0,1\}$ and any $\gamma_d \in \mathcal{P}^{\mathcal{S}}$, we have that $E[Y(d)] = \int_{\mathcal{S}} y d\gamma_d(s) = y$. Since experimental data only affect the feasible γ_d, the result follows. □

Lemma 11. *Let Assumptions RA, EV, PS and BS hold. Suppose $\mathcal{S}$ is a finite set.*

i) Suppose that Assumption LIV holds. Then for $s \in \mathcal{S}$:

$$\begin{aligned}L_\gamma(s) = \Bigg(&\min_{s' \geq s} E_O[Y|S=s', D=0]\frac{P_O(S=s', D=0)}{\gamma_0(s')} + 1 - \frac{P_O(S=s', D=0)}{\gamma_0(s')}, \\ &\max_{s' \leq s} E_O[Y|S=s', D=1]\frac{P_O(S=s', D=1)}{\gamma_1(s')}\Bigg), \\ U_\gamma(s) = \Bigg(&\max_{s' \leq s} E_O[Y|S=s', D=0]\frac{P_O(S=s', D=0)}{\gamma_0(s')}, \\ &\min_{s' \geq s} E_O[Y|S=s', D=1]\frac{P_O(S=s', D=1)}{\gamma_1(s')} + 1 - \frac{P_O(S=s', D=1)}{\gamma_1(s')}\Bigg),\end{aligned}$$

ii) Suppose that Assumption TI holds. Then $L_{\gamma'} = (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ *and* $U_{\gamma'} = (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$ *for:*

$$\begin{aligned} m(s)^{TI,L,\gamma'} &= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] \\ m(s)^{TI,U,\gamma'} &= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)], \end{aligned} \tag{117}$$

where:

$$\begin{aligned} m(s)^{L,\gamma'} &= \min_{d \in \{0,1\}} E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma_d'(s)}, \\ m(s)^{U,\gamma'} &= \max_{d \in \{0,1\}} E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma_d'(s)} + 1 - \frac{P_O(S=s, D=d)}{\gamma_d'(s)}. \end{aligned} \tag{118}$$

Proof. i)

Fix any γ' such that there exists $(m, \gamma') \in \mathcal{H}(m, \gamma)$. Then $\mathcal{H}(m|\gamma') \neq \emptyset$. By Assumption BS, $h_{co(\mathcal{Y})}(-1) = 0$ and $h_{co(\mathcal{Y})}(1) = 1$. By Proposition 2, $\forall s \in \mathcal{S}$ restrictions imposed by data on $m_d(s)$ can then be equivalently stated for $d \in \{0, 1\}$ as:

$$\begin{aligned} m_d(s) \in \Bigg[& E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma_d'(s)}, \\ & E_O[Y|S=s, D=d] \frac{P_O(S=s, D=d)}{\gamma_d'(s)} + 1 - \frac{P_O(S=s, D=d)}{\gamma_d'(s)} \Bigg]. \end{aligned} \tag{119}$$

By Manski and Pepper (2000, Proposition 1) under Assumption LIV the sharp bound on $m_d(s)$ is:

$$\begin{aligned} m_d(s) \geq m_d(s)^{LIV,L,\gamma'} &:= \sup_{s' \leq s} E_O[Y|S=s', D=d] \frac{P_O(S=s', D=d)}{\gamma_d'(s')} \\ m_d(s) \leq m_d(s)^{LIV,U,\gamma'} &:= \inf_{s' \geq s} E_O[Y|S=s', D=d] \frac{P_O(S=s', D=d)}{\gamma_d'(s')} + 1 - \frac{P_O(S=s', D=d)}{\gamma_d'(s')}. \end{aligned} \tag{120}$$

First, note that both $m_d^{LIV,L,\gamma'}$ and $m_d^{LIV,U,\gamma'}$ are non-decreasing in s by definition for all $d \in \{0, 1\}$. Thus, $L_{\gamma'} := (m_0^{LIV,U,\gamma'}, m_1^{LIV,L,\gamma'}) \in \mathcal{M}^A$ and $U_{\gamma'} := (m_0^{LIV,L,\gamma'}, m_1^{LIV,U,\gamma'}) \in \mathcal{M}^A$. Hence $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m, \gamma)$. Since γ' was arbitrary, $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m, \gamma)$ for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$. Then, observe that T is non-decreasing in $m_1(s)$ and non-increasing in $m_0(s)$ for each $s \in \mathcal{S}$. Therefore, $\forall m \in \mathcal{H}(m|\gamma')$ $T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to T. Similarly, $\forall m \in \mathcal{H}(m|\gamma')$ $T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to T. Since $|\mathcal{S}|$ is a compact set, infima and

suprema may be replaced by minima and maxima.

ii)

As in proof of *i)*, fix any γ' such that there exists $(m, \gamma') \in \mathcal{H}(m, \gamma)$, so $\mathcal{H}(m|\gamma') \neq \emptyset$. Assumption TI maintains that $m_1 = m_0$. Then write for any $s \in \mathcal{S}$ and $d \in \{0, 1\}$:

$$T(m, \gamma') = \int_{\mathcal{S}} m_1(s) d\gamma_1'(s) - \int_{\mathcal{S}} m_0(s) d\gamma_0'(s) = \int_{\mathcal{S}} m_d(s)(d\gamma_1'(s) - d\gamma_0'(s)). \tag{121}$$

Define:

$$\begin{aligned} m(s)^{L,\gamma'} &:= \min_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)}, \\ m(s)^{U,\gamma'} &:= \max_{d \in \{0,1\}} E_O[Y|S = s, D = d] \frac{P_O(S = s, D = d)}{\gamma_d'(s)} + 1 - \frac{P_O(S = s, D = d)}{\gamma_d'(s)}. \end{aligned} \tag{122}$$

Next let for any $s \in \mathcal{S}$:

$$\begin{aligned} m(s)^{TI,L,\gamma'} &:= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] \\ m(s)^{TI,U,\gamma'} &:= m(s)^{L,\gamma'} \mathbb{1}[\gamma_1'(s) < \gamma_0'(s)] + m(s)^{U,\gamma'} \mathbb{1}[\gamma_1'(s) \geq \gamma_0'(s)] \end{aligned} \tag{123}$$

and $L_{\gamma'} := (m^{TI,L,\gamma'}, m^{TI,L,\gamma'})$ and $U_{\gamma'} := (m^{TI,U,\gamma'}, m^{TI,U,\gamma'})$.

By Proposition 2, it is immediate that for $\mathcal{H}(m|\gamma') = \{m \in \mathcal{M} : m_1 = m_0, \ \forall d \in \{0, 1\}, \ \forall s \in \mathcal{S}, \ m_d(s) \geq m(s)^{L,\gamma'}, \ m_d(s) \leq m_d(s)^{U,\gamma'}\}$. Hence $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$. Since γ' was arbitrary, $(L_{\gamma'}, \gamma'), (U_{\gamma'}, \gamma') \in \mathcal{H}(m|\gamma')$ for any $\gamma' \in \mathcal{H}(\gamma)$. Therefore, $L_{\gamma'}$ and $U_{\gamma'}$ are selectors of $\mathcal{H}(m|\cdot)$.

Then observe that by (121), $\forall m \in \mathcal{H}(m|\gamma') \ T(L_{\gamma'}, \gamma') \leq T(m, \gamma')$, so $L_{\gamma'}$ is a minimal selector with respect to T. Similarly, $\forall m \in \mathcal{H}(m|\gamma') \ T(U_{\gamma'}, \gamma') \geq T(m, \gamma')$, so $U_{\gamma'}$ is a maximal selector with respect to T. □

References

Aizer, Anna, Nancy Early, Shari Eli, Guido Imbens, Keyoung Lee, Adriana Lleras-Muney, and Alexander Strand. 2024. "The Lifetime Impacts of the New Deal's Youth Employment Program." *The Quarterly Journal of Economics.*

Al-Khayyal, Faiz A. 1992. "Generalized bilinear programming: Part I. Models, applications and linear programming relaxation." *European Journal of Operational Research* 60 (3): 306–314.

Artstein, Zvi. 1983. "Distributions of random sets and random selections." *Israel Journal of Mathematics* 46:313–324.

Athey, Susan, Raj Chetty, and Guido Imbens. 2020. "Combining experimental and observational data to estimate treatment effects on long term outcomes." *arXiv preprint arXiv:2006.09676.*

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2024. *Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.* Technical report.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago. 2012. "Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa." *The Review of Economic Studies* 79 (1): 37–66.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2012. "Partial identification using random set theory." *Journal of Econometrics* 166 (1): 17–32.

———. 2011. "Sharp identification regions in models with convex moment predictions." *Econometrica* 79 (6): 1785–1821.

Billingsley, P. 1995. "Probability and measure." *Wiley series in probability and mathematical statistics.*

Chen, Jiafeng, and David M Ritzwoller. 2023. "Semiparametric estimation of long-term treatment effects." *Journal of Econometrics* 237 (2): 105545.

Chen, Xuan, Carlos A Flores, and Alfonso Flores-Lagunes. 2018. "Going beyond LATE: bounding average treatment effects of Job Corps training." *Journal of Human Resources* 53 (4): 1050–1099.

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

Currie, Janet, and Douglas Almond. 2011. "Human capital development before age five." In *Handbook of labor economics,* 4:1315–1486. Elsevier.

Deaton, Angus S. 2009. *Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development.* Technical report. National bureau of economic research.

Galichon, Alfred. 2018. *Optimal transport methods in economics.* Princeton University Press.

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

García, Jorge Luis, James J Heckman, Duncan Ermini Leaf, and María José Prados. 2020. "Quantifying the life-cycle benefits of an influential early-childhood program." *Journal of Political Economy* 128 (7): 2502–2541.

Ghassami, AmirEmad, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. 2022. "Combining experimental and observational data for identification and estimation of long-term causal effects." *arXiv preprint arXiv:2201.10743.*

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. "Top challenges from the first practical online controlled experiments summit." *ACM SIGKDD Explorations Newsletter* 21 (1): 20–35.

Han, Sukjin, and Hiroaki Kaido. 2024. "Set-Valued Control Functions." *arXiv preprint arXiv:2403.00347.*

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review* 103 (6): 2052–2086.

Heckman, James J, and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156 (1): 27–37.

Heckman, James J, and Edward J Vytlacil. 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.

Hoynes, Hilary W, and Diane Whitmore Schanzenbach. 2018. *Safety net investments in children.* Technical report. National Bureau of Economic Research.

Hu, Wenjie, Xiaohua Zhou, and Peng Wu. 2022. "Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data." *arXiv preprint arXiv:2208.10163.*

Huber, Martin, Lukas Laffers, and Giovanni Mellace. 2015. "Sharp IV bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance." *Journal of Applied Econometrics* 32 (1): 56–79.

Imbens, Guido, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2024. "Long-term causal inference under persistent confounding via data combination." *arXiv preprint arXiv:2202.07234.*

Imbens, Guido W. 2010. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic literature* 48 (2): 399–423.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

Kamat, Vishal. 2024. "Identifying the effects of a program offer with an application to head start." *Journal of Econometrics* 240 (1): 105679.

Kitagawa, Toru. 2015. "A test for instrument validity." *Econometrica* 83 (5): 2043–2063.

Kline, Patrick, and Christopher R Walters. 2016. "Evaluating public programs with close substitutes: The case of Head Start." *The Quarterly Journal of Economics* 131 (4): 1795–1848.

Li, Shoumei, and Yukio Ogura. 1998. "Convergence of set valued sub-and supermartingales in the Kuratowski-Mosco sense." *Annals of probability,* 1384–1402.

Manski, Charles F. 1997. "Monotone treatment response." *Econometrica: Journal of the Econometric Society,* 1311–1334.

———. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review* 80 (2): 319–323.

Manski, Charles F, and John V Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010.

———. 2009. "More on monotone instrumental variables." *The Econometrics Journal* 12 (suppl_1): S200–S216.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86 (5): 1589–1619.

Molchanov, Ilya. 2017. *Theory of Random Sets.* 2nd ed. Vol. 87. Probability Theory and Stochastic Modelling. Springer.

Molchanov, Ilya, and Francesca Molinari. 2014. "Applications of random set theory in econometrics." *Annu. Rev. Econ.* 6 (1): 229–251.

———. 2018. *Random Sets in Econometrics.* Vol. 60. Cambridge University Press.

Moon, Sarah. 2024. "Partial Identification of Individual-Level Parameters Using Aggregate Data in a Nonparametric Binary Outcome Model." *arXiv preprint arXiv:2403.07236.*

Mourifié, Ismael, and Yuanyuan Wan. 2017. "Testing local average treatment effect assumptions." *Review of Economics and Statistics* 99 (2): 305–313.

Obradović, Filip. 2024. "A test for external validity in data combination." *Working paper.*

Pages, Remy, Dylan J Lukes, Drew H Bailey, and Greg J Duncan. 2020. "Elusive longer-run impacts of head start: Replications within and across cohorts." *Educational Evaluation and Policy Analysis* 42 (4): 471–492.

Park, Yechan, and Yuya Sasaki. 2024a. "A Bracketing Relationship for Long-Term Policy Evaluation with Combined Experimental and Observational Data." *arXiv preprint arXiv:2401.12050*.

———. 2024b. *The Informativeness of Combined Experimental and Observational Data under Dynamic Selection*. arXiv: 2403.16177 [econ.EM].

Ponomarev, Kirill. 2024. *Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions*. http://kponomarev.github.io/files_on_website/sharp%20inequalities.pdf.

Prentice, Ross L. 1989. "Surrogate endpoints in clinical trials: definition and operational criteria." *Statistics in medicine* 8 (4): 431–440.

Rockafellar, Ralph Tyrell. 1970. *Convex Analysis*. Princeton: Princeton University Press. ISBN: 9781400873173. https://doi.org/doi:10.1515/9781400873173. https://doi.org/10.1515/9781400873173.

Russell, Thomas M. 2021. "Sharp bounds on functionals of the joint distribution in the analysis of treatment effects." *Journal of Business & Economic Statistics* 39 (2): 532–546.

Schaefer, Helmut H., and M. P. Wolff. 1999. *Topological Vector Spaces*. Springer.

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

Todd, Petra E, and Kenneth I Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American economic review* 96 (5): 1384–1417.

———. 2023. "The best of both worlds: combining randomized controlled trials with structural modeling." *Journal of Economic Literature* 61 (1): 41–85.

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

Treves, François. 2016. *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25*. Vol. 25. Elsevier.

Van Goffrier, Graham, Lucas Maystre, and Ciarán Mark Gilligan-Lee. 2023. "Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding." In *Conference on Causal Learning and Reasoning*, 791–813. PMLR.

Vikström, Johan, Geert Ridder, and Martin Weidner. 2018. "Bounds on treatment effects on transitions." *Journal of Econometrics* 205 (2): 448–469.

Villani, Cédric, et al. 2009. *Optimal transport: old and new.* Vol. 338. Springer.

Willard, Stephen. 2004. *General topology.* Courier Corporation.

Notation

I denote random variables and vectors using capital letters (e.g. Y), and their laws by P (e.g. P_Y). I denote random sets with boldface letters (e.g. $\mathbf{Y}$) and $\mathbb{E}(\mathbf{Y}|X)$ is used for the conditional Aumann expectation of a random set $\mathbf{Y}$ given a sigma-algebra generated by a random variable X. I use $\overset{d}{=}$ to denote that a random element has a certain law, or an equivalent distribution-determining functional. (e.g. $Y \overset{d}{=} P_Y$) I also use $\overset{d}{=}$ to denote equality of distribution of two random elements. (e.g. $Y \overset{d}{=} Y'$) Supports of random variables are denoted by corresponding script letters (e.g. $\mathcal{Y}$) and the support is defined as the smallest closed set containing the random variable a.s. I also use capital letters to denote sets (usually A, B and K), and denote by $\mathcal{C}(A)$, $\mathcal{B}(A)$ the families of all closed and Borel subsets of the set A, respectively. Let $co(A)$ be the closed convex hull of the set A. The sharp identified sets for a generic parameter θ is denoted by $\mathcal{H}(\theta)$, and for the distribution function (i.e. the law or equivalently the CDF) of a generic random vector Y by $\Gamma(Y)$ and $\mathcal{H}(P_Y)$. Finally, let the set of distribution functions of random variables with support $\mathcal{Y}$ be $\mathcal{P}^{\mathcal{Y}}$.