Inference for Treatment Effects Conditional on Generalized Principal Strata using Instrumental Variables *

Yuehao Bai
Department of Economics
University of Southern California
yuehao.bai@usc.edu

Shunzhuang Huang
Booth School of Business
University of Chicago
shunzhuang.huang@chicagobooth.edu

Sarah Moon
Department of Economics
Massachusetts Institute of Technology
sarahmn@mit.edu

Andres Santos
Department of Economics
University of California–Los Angeles
andres@econ.ucla.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Edward J. Vytlacil
Department of Economics
Yale University
edward.vytlacil@yale.edu

November 15, 2025

Abstract

We propose a general approach to inference for a broad class of models that arise in the analysis of treatment effects with discrete-valued treatments and instruments and a general-valued outcome. In addition to instrument exogeneity, the main substantive assumption in our class of models rules out certain response types by assuming that they occur with probability zero. Here, the response type refers to the vector of potential outcomes and potential treatments, and we refer to a set of possible values for the response type as a generalized principal stratum. Through a series of examples, we show that this framework encompasses a wide variety of assumptions that have been considered in the previous literature. Our framework allows inference on any treatment effect parameter that can be expressed as the expectation of a function of the response type conditional on a generalized principal stratum. We develop methods for inference on such parameters under these assumptions, as well as methods for testing the validity of the assumptions themselves. A key result of our analysis is a characterization of the identified set for such parameters under these assumptions and the testable restrictions for the assumptions themselves in terms of existence of a non-negative solution to linear systems of equations with a special structure. We propose methods for inference exploiting this special structure and recent results in Fang et al. (2023).

KEYWORDS: Multi-valued Treatments, Model Specification, Model Validity, Randomized Controlled Trial, Principal Strata, Instrumental Variables, Partial Identification.

*We thank Jing Cheng, Jiaying Gu, Sukjin Han, Nathan Kallus, Soonwoo Kwon, Sokbae Lee, Derek Neal, Kirill Ponomarev, Vitor Possebom, Bernard Salanié and Panos Toulis for helpful comments. Shaikh acknowledges support from NSF Grant SES-2419008; Moon acknowledges support from NSF Grant No. 2141064; Vytlacil acknowledges financial support from the Tobin Center for Economic Policy at Yale.

1 Introduction

We propose a general approach to inference for a broad class of models that arise in the analysis of treatment effects with discrete-valued treatments and instruments and a general-valued outcome. In addition to instrument exogeneity, the main substantive restriction imposed in our analysis is that certain values for the response types occur with probability zero. Here, the response type refers to the vector of potential outcomes and potential treatments, and we refer to a set of possible values for the response type as a generalized principal stratum. Through a series of examples, we show that our framework encompasses a wide variety of assumptions that have been considered in the previous literature, including the following: (i) restrictions in the analysis of randomized controlled trials (RCTs) under noncompliance considered, e.g., in Imbens and Angrist (1994) and Cheng and Small (2006); (ii) generalizations of these restrictions considered in Bai et al. (2024); (iii) revealed preference-type restrictions on response types considered in Kirkeboen et al. (2016), Kline and Walters (2016), and Heckman and Pinto (2018); and (iv) restrictions on the ordering of potential treatments or on the ordering of potential outcomes considered in Manski (1997), Manski and Pepper (1998), and Machado et al. (2019).[1]

Our framework allows inference on any treatment effect parameter that can be expressed as the expectation of a function of the response type conditional on a generalized principal stratum. In this way, our framework accommodates many parameters that have been considered previously in the literature, including both average and distributional treatment effect parameters, such as the probability of being strictly helped and the probability of not being hurt by the treatment. It further accommodates versions of these parameters conditional on sets of possible values of potential treatments, such as the local average treatment effect in Imbens and Angrist (1994), average effects conditional on principal strata in Frangakis and Rubin (2002), and average effects conditional on sets of possible values of potential outcomes, such as the parameters considered in Heckman et al. (1997) and Heckman and Smith (1998). By contrast, each of the papers cited in the preceding paragraph tailors their method to specific examples of such treatment effect parameters. Furthermore, because the types of restrictions considered in these papers are typically insufficient for identification of many parameters of interest, some of these papers focus specifically on those treatment effect parameters that are identified from the two-stage least squares estimand. A characterization of the parameters that are identified by such restrictions is given by Navjeevan et al. (2023) and Goff (2025).

[1] In the context of mediation analysis, Kwon and Roth (2024) show that "full mediation" is equivalent to certain assumptions like those we consider.

A key result in our analysis is a characterization of the identified set for such parameters under these assumptions and the testable restrictions for the assumptions themselves in terms of existence of a non-negative solution to linear systems of equations with a special structure. Using this result, we build on results in Fang et al. (2023) to develop a test for the null hypothesis that a pre-specified value for the parameter of interest lies in the identified set as well as the null hypothesis that the testable restrictions are satisfied. Importantly, the resulting tests remain well behaved even in "high-dimensional" settings, meaning it is uniformly consistent in level over a large class of distributions satisfying weak assumptions and permitting, among other things, the support of the treatment and instrument or the support of the response types to be "large" relative to the sample size. Through test inversion, we may also construct confidence regions for such parameters that are uniformly consistent in level. We highlight that while our test builds on Fang et al. (2023), it leverages the structure of our problem to obtain novel strong approximations that significantly improve on the coupling rates of Fang et al. (2023). These coupling results for high-dimensional vectors of estimates of conditional probabilities and their bootstrap counterparts may be of independent interest.

Other methods for inference have been proposed in the prior literature for some special cases of our framework. For inference on certain treatment effect parameters, see, e.g., Cheng and Small (2006), Bhattacharya et al. (2008, 2012), and Machado et al. (2019); for inference on the validity of the assumptions, see, e.g., Kitagawa (2015), Heckman and Pinto (2018), Sun (2023), and Bai et al. (2025). These approaches rely upon closed-form expressions for the identified set for the parameter of interest or testable inequalities, which must be proved or computed on a case-by-case basis. When such expressions need to be computed, the methods are feasible only when the support of response types is small. In contrast, our approach does not rely on such closed-form expressions and remains computationally feasible even in high-dimensional settings. A key strength of our approach is its broad applicability, including in settings for which no existing methods apply.

The remainder of the paper is organized as follows. Section 2 introduces our formal setup and notation. In Section 3, we provide several examples of parameters and restrictions previously considered in the literature that are nested by our framework. We present our inference method when the support of the outcome variable is restricted to be finite in Section 4, and without such a restriction in Section 5. We illustrate our method through a simulation study in Section 6. Proofs of all results can be found in Appendix A.

2 Setup and Notation

Denote by $Y \in \mathcal{Y}$ an outcome, by $D \in \mathcal{D}$ a discrete valued endogenous regressor (i.e., the treatment), and by $Z \in \mathcal{Z}$ a discrete valued instrumental variable. To rule out degenerate cases, we assume throughout that $2 \leq |\mathcal{Y}|$, $2 \leq |\mathcal{D}| < \infty$, and $2 \leq |\mathcal{Z}| < \infty$. We emphasize that neither $\mathcal{D}$ and $\mathcal{Z}$ needs to be a subset of the real line; in this way, our framework accommodates vector-valued treatments and instruments provided they only take a finite number of values. In what follows, we will consider both settings in which $|\mathcal{Y}|$ is finite and those in which it is not. Further denote by $Y(d) \in \mathcal{Y}$ the potential outcome if $D = d \in \mathcal{D}$ and by $D(z) \in \mathcal{D}$ the potential treatment if $Z = z \in \mathcal{Z}$. As usual, we assume that

$$Y = \sum_{d \in \mathcal{D}} Y(d) I\{D = d\} \quad \text{and} \quad D = \sum_{z \in \mathcal{Z}} D(z) I\{Z = z\} \ . \tag{1}$$

In our discussion below, it will be convenient to define $R_o \equiv (Y(d) : d \in \mathcal{D})$ and $R_t \equiv (D(z) : z \in \mathcal{Z})$, and set $R \equiv (R_o, R_t)$. Following Heckman and Pinto (2018), we will refer to R_t as the "treatment response type." By analogy with this terminology, we will also refer to R_o as the "outcome response type" and to R as simply the "response type." We let Q denote the distribution of (R_o, R_t, Z) and note that by (1) we have

$$(Y, D, Z) = T(R_o, R_t, Z) \ ,$$

for a transformation T implicitly defined through (1). Letting P denote the distribution of (Y, D, Z), we therefore have that $P = QT^{-1}$.

Below we will require that $Q \in \mathbf{Q}$, where $\mathbf{Q}$ is a class of distributions satisfying assumptions that we will specify. Different specifications of $\mathbf{Q}$ represent different assumptions that we impose on the distribution of potential outcomes and potential treatments. In this sense, $\mathbf{Q}$ may be viewed as a model for potential outcomes and potential treatments.

Given a distribution P of (Y, D, Z) and a model $\mathbf{Q}$, we define the set of $Q \in \mathbf{Q}$ that can rationalize P as

$$\mathbf{Q}_0(P, \mathbf{Q}) = \{Q \in \mathbf{Q} : P = QT^{-1}\} \ .$$

We say $\mathbf{Q}$ is consistent with P if and only if $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. For every model $\mathbf{Q}$ considered in this paper, every $Q \in \mathbf{Q}$ is assumed to satisfy the restriction:

Assumption 2.1 (Instrument Exogeneity)**.** $R \perp\!\!\!\perp Z$ under Q.

Our remaining restrictions on $\mathbf{Q}$ will be formulated in terms of restrictions on possible values of the response type R. These restrictions will be expressed by specifying a set $\mathcal{R} \subseteq \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$ characterizing the possible values of R. In Section 3, we provide several different choices of $\mathcal{R}$ that have been previously considered in the literature. For a given choice of $\mathcal{R}$, we will impose the following on every $Q \in \mathbf{Q}$:

Assumption 2.2 (Response Type Restrictions)**.** $Q\{R \in \mathcal{R}\} = 1$.

Following Frangakis and Rubin (2002), sets of the form $\{R_t = r_t\}$ for r_t a possible value of R_t are referred to as "principal strata." By analogy with this terminology, we will refer to sets of the form $\{R \in \mathcal{R}'\}$ for $\mathcal{R}' \subseteq \mathcal{R}$ as "generalized principal strata."

Finally, we define our parameters of interest. The parameters we consider can be written as

$$\theta(Q) \equiv E_Q[g(R) \mid R \in \mathcal{R}'] \tag{2}$$

for different choices of function $g : \mathcal{R} \to \mathbf{R}$ and generalized principal strata $\mathcal{R}' \subseteq \mathcal{R}$. Note that for $\theta(Q)$ in (2) to be well defined we require Q to satisfy $Q\{R \in \mathcal{R}'\} > 0$. A wide variety of parameters can be accommodated in this way. In Section 3, we will show specific parameters that have been considered previously in the literature have the structure in (2). We note now, however, that natural choices of g correspond to the effect of one treatment versus another (i.e., $g(R) = Y(d) - Y(d')$ for $d \in \mathcal{D}$ and $d' \in \mathcal{D}$) and the probability that one treatment leads to a larger outcome than another treatment (i.e., $g(R) = I\{Y(d) > Y(d')\}$ for $d \in \mathcal{D}$ and $d' \in \mathcal{D}$). We further note that when $\mathcal{R}' = \mathcal{R}$, $\theta(Q)$ defined in (2) simplifies to $E_Q[g(R)]$.

For a given distribution P and model $\mathbf{Q}$, note that the identified set for $\theta(Q)$ under P relative to $\mathbf{Q}$ is

$$\Theta_0(P, \mathbf{Q}) \equiv \{\theta(Q) : Q \in \mathbf{Q}_0(P, \mathbf{Q}) \text{ and } Q\{R \in \mathcal{R}'\} > 0\} . \tag{3}$$

The set $\Theta_0(P, \mathbf{Q})$ is nonempty whenever there exists at least one $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ with $Q\{R \in \mathcal{R}'\} > 0$. By construction, this set is "sharp" in the sense that for any value in the set there exists a distribution Q that is consistent with P, satisfies the restrictions of the model, and for which $\theta(Q)$ equals the prescribed value.

3 Examples

In this section, we show how to accommodate several examples from the previous literature in our framework. Our discussion focuses in particular on Assumption 2.1 and the specification of $\mathcal{R}$ in Assumption 2.2, but, where the cited literature has emphasized specific parameters of interest, we additionally describe how those parameters of interest can be expressed as (2) for suitable choices of g and $\mathcal{R}'$.

Example 3.1 (*RCT with one-sided noncompliance*)**.** Consider a multi-arm randomized controlled trial (RCT) with noncompliance, where $\mathcal{D} = \mathcal{Z} = \{0, \ldots, K\}$, $Z = d$ denotes random assignment to treatment d, and $D(d) = d$ denotes that the subject would comply with assignment if assigned to treatment d. In this example, Q satisfies Assumption 2.1 because Z is randomly assigned. Suppose noncompliance to the assignment is one-sided in the sense that one can always take the control $d = 0$, but, for any other treatment $d \neq 0$, one can only take that treatment if assigned to it. This restriction can be expressed in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), \ldots, y(K), d(0), \ldots, d(K)) : d(j) \in \{0, j\}$ for all $j \in \mathcal{D}\}$. With the notable exception of Cheng and Small (2006), discussed below, analyses of causal parameters that condition on generalized principal strata in this context follow Imbens and Angrist (1994) in using the Wald estimand to identify local average treatment effect (LATE) parameters of the form $E_Q[Y(j) - Y(0) \mid D(j) = 1]$ for $j \in \{1, \ldots, K\}$. Note that identification of such parameters does not allow comparison of $Y(j)$ to $Y(k)$ for $j, k \neq 0$ for any subgroup of subjects. Our framework nests these identified parameters as well as partially identified parameters including the relative treatment effectiveness for any subgroup defined by a generalized principal stratum. ■

Example 3.2. Cheng and Small (2006) study the special case of Example 3.1 in which $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$. Their "Monotonicity I" assumption corresponds to Assumption 2.2 with $\mathcal{R}$ defined as in Example 3.1. They further consider imposing the restriction that $Q\{D(1) = 1 \mid D(2) = 2\} = 1$, i.e., that subjects who would comply with assignment to treatment 2 would also comply with assignment to treatment 1. They argue that this assumption is plausible in contexts where the "cost" of compliance with treatment 1 is lower than that with treatment 2. The combination of these two restrictions can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), y(1), y(2), d(0), d(1), d(2)) : (d(0), d(1), d(2)) \in \{(0, 0, 0), (0, 1, 0), (0, 1, 2)\}\}$. In their application, Cheng and Small (2006) focus on the following parameters: (i) $E_Q[Y(j) - Y(0) \mid (D(0), D(1), D(2)) = (0, 1, 2)]$ for $j \in \{1, 2\}$; and (ii) $Q\{(D(0), D(1), D(2)) = r_t\}$ for different values of r_t. Each of these parameters can be expressed in the form of (2) for appropriate choices of g and $\mathcal{R}'$. For example, the parameter in (i) equals $E_Q[g(R) \mid R \in \mathcal{R}']$ for $g(R) = Y(j) - Y(0)$ and $\mathcal{R}' = \{(y(0), y(1), y(2), d(0), d(1), d(2)) \in \mathcal{R} : (d(0), d(1), d(2)) = (0, 1, 2)\}$. Our framework nests their parameters within their context, while also allowing for inference on the corresponding parameters for trials with more than three treatment arms, for which their approach becomes computationally infeasible. ■

Example 3.3 (*Encouragement design*)**.** Consider a multi-arm RCT with possibly two-sided noncompliance, where $\mathcal{D} = \mathcal{Z} = \{0, \ldots, K\}$, $Z = d$ denotes random assignment to treatment d, and $D(d) = d$ denotes that the subject would comply with assignment if assigned to treatment d. More generally, not necessarily in the context of an RCT, one can interpret $Z = d$ as random encouragement to treatment d and interpret $D(d) = d$ as the subject would take treatment d if encouraged to do so. In this example, Q satisfies Assumption 2.1 because Z is randomly assigned. Bai et al. (2024) generalize the "no-defier" restriction of Imbens and Angrist (1994) to

$$Q\{D(d) \neq d,\ D(d') = d \text{ for some } d' \neq d\} = 0 \ , \tag{4}$$

i.e., a subject that would not take treatment d if assigned to (encouraged to take) d but would also not take d if assigned (encouraged) to some other treatment $d' \neq d$. This restriction can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), \ldots, y(K), d(0), \ldots, d(K)) : d(j) \neq j \ \Rightarrow d(k) \neq j \ \forall \ j, k \in \mathcal{D}\}$. Bai et al. (2024) derive the identified set on unconditional average treatment effects in this context under (4). While the two-stage least squares (TSLS) estimand in this context generally does not correspond to any well-defined causal parameter when $K \geq 2$, Bhuller and Sigstad (2024) show that under strong, additional assumptions, it identifies a particular weighted average of strata-specific causal effects that may not be of a priori interest. In contrast, our framework allows inference on a broad class of parameters of substantive interest including those that condition on generalized principal strata and that need not coincide with the TSLS estimand. ■

Example 3.4 (*RCT with close substitute*)**.** Kline and Walters (2016) consider an RCT with a "close substitute" to study the effects of preschooling on educational outcomes. In their setting, $D \in \mathcal{D} = \{c, h, n\}$, where $D = c$ denotes home care (no preschool), $D = h$ denotes a preschool program called Head Start, and $D = n$ denotes preschools other than Head Start, i.e., the close substitute. Let $Z \in \mathcal{Z} = \{0, 1\}$ denote an indicator variable for a randomized offer to attend Head Start. Assumption 2.1 holds because Z is randomly assigned. In evaluating the cost-effectiveness of Head Start, Kline and Walters (2016) impose the restriction

$$Q\{D(1) = h \mid D(0) \neq D(1)\} = 1 \ . \tag{5}$$

The condition in (5) states that if a family's schooling choice changes upon receiving a Head Start offer, then they must choose Head Start when receiving the offer. In other words, it cannot be the case that upon receiving a Head Start offer, a family switches from no preschool to preschools other than Head Start, or the other way around. The restriction can be formulated in terms of Assumption 2.2 with

$$\mathcal{R} = \{(y(c), y(h), y(n), d(0), d(1)) : (d(0), d(1)) \in \{(c, c), (h, h), (n, n), (c, h), (n, h)\}\} .$$

Kline and Walters (2016) show that the Wald estimand identifies a weighted combination of "sub-LATEs":

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{P\{D = h \mid Z = 1\} - P\{D = h \mid Z = 0\}} = S_c \text{SubLATE}_{ch} + (1 - S_c)\text{SubLATE}_{nh} ,$$

where

$$\begin{aligned} \text{SubLATE}_{ch} &= E_Q[Y(h) - Y(c) \mid D(1) = h, D(0) = c] \\ \text{SubLATE}_{nh} &= E_Q[Y(h) - Y(n) \mid D(1) = h, D(0) = n] \end{aligned}$$

and $S_c = Q\{D(1) = h, D(0) = c \mid D(1) = h, D(0) \neq h\}$. Here, S_c and the sub-LATEs can all be written as (2) for appropriate choices of g and $\mathcal{R}'$. Kline and Walters (2016) show that these separate sub-LATE parameters are of substantive interest, and identify them under additional assumptions including imposing a multinomial normal selection model. Our framework allows inference on these sub-LATE parameters without imposing such additional assumptions sufficient to identify them. ■

Example 3.5. Kirkeboen et al. (2016) study the effects of college field of study on earnings. In their setting, $\mathcal{D} = \{0, 1, 2\}$ represents three fields of study, ordered by their (soft) admission cutoffs from lowest to highest. The instrument is $Z \in \{0, 1, 2\}$, where $Z = 1$ when the student crosses the (soft) admission cutoff for field 1 but not for field 2, $Z = 2$ when the student crosses the (soft) admission cutoff for field 2, and $Z = 0$ otherwise. The authors assume that Z is exogenous in the sense that Q satisfies Assumption 2.1, and they impose:

$$Q\{D(1) = 1 \mid D(0) = 1\} = 1 , \tag{6}$$

$$Q\{D(2) = 2 \mid D(0) = 2\} = 1 . \tag{7}$$

The monotonicity conditions in (6)–(7) require that crossing the cutoff for field 1 or 2 weakly encourages students toward that field. They further impose the following "irrelevance" conditions:

$$Q\{I\{D(1) = 2\} = I\{D(0) = 2\} \mid D(0) \neq 1, D(1) \neq 1\} = 1 , \tag{8}$$

$$Q\{I\{D(2) = 1\} = I\{D(0) = 1\} \mid D(0) \neq 2, D(2) \neq 2\} = 1 . \tag{9}$$

The condition in (8) states that if crossing the cutoff for field 1 does not cause a student to switch to field 1, then it also does not cause them to switch to or away from field 2. A similar interpretation applies to (9). The restrictions in (6)–(9) can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), y(1), y(2), d(0), d(1), d(2)) : (d(0), d(1), d(2)) \in \{(0, 0, 0), (0, 0, 2), (0, 1, 0), (0, 1, 2), (1, 1, 1), (1, 1, 2), (2, 1, 2), (2, 2, 2)\}\}$. Kirkeboen et al. (2016) regard the irrelevance conditions (8)–(9) as strong but, under them, show that the TSLS estimands identify $E_Q[Y(2) - Y(0) \mid D(1) = 1, D(0) = 0]$ and $E_Q[Y(2) - Y(0) \mid D(2) = 2, D(0) = 0]$. Under their assumptions (6)–(9), our analysis allows one to analyze additional parameters that do not correspond to the TSLS estimands, including the average relative effects of field 1 versus field 2 for any given generalized principal strata. Our framework further allows one to investigate what can be learned about these parameters without imposing (8)–(9). For further discussion of this example and Example 3.4, see Lee and Salanié (2023) and Bai et al. (2025). ■

Example 3.6 (*Restrictions from WARP*)**.** Heckman and Pinto (2018) consider a setting in which there is a voucher Z that subsidizes in different ways that we specify below the purchase of three different cars, that we denote by A, B and C. They further assume that the voucher is randomly assigned, so that Assumption 2.1 holds. The treatment D corresponds to the purchase of the different cars; let $D = A$ correspond to the purchase of car A, $D = B$ correspond to the purchase of car B, and $D = C$ correspond to the purchase of car C. In this setting, Heckman and Pinto (2018) consider a series of examples in which they use the Weak Axiom of Revealed Preference (WARP) to restrict treatment response types, each of which can be formulated in terms of Assumption 2.2 for appropriate choice of $\mathcal{R}$.

(i) In their leading example, $Z = 0$ corresponds to no voucher, $Z = 1$ corresponds to a voucher that subsidizes the purchase of car A, and $Z = 2$ corresponds to a voucher that subsidizes the purchase of either B or C. WARP generates the restriction in Table III of Heckman and Pinto (2018). These restrictions can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(A), y(B), y(C), d(0), d(1), d(2)) : (d(0), d(1), d(2)) \in \{(A, A, A), (A, A, B), (A, A, C), (B, A, B), (B, B, B), (C, A, C), (C, C, C)\}\}$.

(ii) In a second example, $Z = 0$ corresponds to no voucher, $Z = 1$ corresponds to a voucher that subsidizes the purchase of B, and $Z = 2$ corresponds to a voucher that subsidizes the purchase of B or C. WARP generates the restriction in Table V of Heckman and Pinto (2018), which can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(A), y(B), y(C), d(0), d(1), d(2)) : (d(0), d(1), d(2)) \in \{(A, A, A), (A, A, C), (A, B, B), (A, B, C), (B, B, B), (C, C, C), (C, B, C)\}\}$.

(iii) Finally, in a third example, $Z = 0$ corresponds to a voucher that subsidizes the purchase of C, $Z = 1$ corresponds to a voucher that subsidizes the purchase of B, and $Z = 2$ corresponds to a voucher that subsidizes the purchase of B or C. WARP generates the restriction in Table VI of Heckman and Pinto (2018), which can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(A), y(B), y(C), d(0), d(1), d(2)) : (d(0), d(1), d(2)) \in \{(A, A, A), (A, B, B), (B, B, B), (C, A, C), (C, B, B), (C, B, C), (C, C, C)\}\}$.

Heckman and Pinto (2018) provide several additional examples that also fit within our framework. They establish conditions for identification of average treatment effect parameters that condition on generalized principal strata in these examples; in comparison, our analysis allows inference for a more general class of parameters that may be only partially identified. ■

Example 3.7 (*Ordered monotonicity with known direction*)**.** Consider a setting in which both the treatment and the instrument admit natural orderings, and write $\mathcal{D} = \{0, \ldots, K_d\}$ and $\mathcal{Z} = \{0, \ldots, K_z\}$. Suppose further that Z is randomly assigned, so Assumption 2.1 holds. In many applications, it is reasonable to assume the monotonicity restriction $Q\{D(j) \geq D(k)\} = 1$ for any pair of instrument values $j, k \in \mathcal{Z}$ with $j \geq k$. For example, Dupas (2014) reports an experiment in which Z represents different randomly assigned subsidy levels for insecticide-treated bed nets, and D denotes the total number of bednets purchased over a two-year period. The monotonicity restriction in this context is that one would purchase at least as many insecticide-treated bed nets with a greater subsidy as with a lower subsidy. This restriction can be expressed in terms of Assumption 2.2 as $\mathcal{R} = \{(y(0), \ldots, y(K_d), d(0), \ldots, d(K_z)) : d(K_z) \geq \ldots \geq d(0)\}$. The results of Angrist and Imbens (1995) imply that the TSLS estimand in this context identifies a specific weighted average of strata-specific causal effects, whereas our analysis allows inference on a broader class of parameters—including partially identified parameters—that need not coincide with the TSLS estimand. ■

Example 3.8 (*Monotone treatment response*)**.** Consider a setting in which the treatment admits a natural ordering and write $\mathcal{D} = \{0, \ldots, K_d\}$. Suppose further that Z is randomly assigned, so Assumption 2.1 holds. Following Manski (1997), it is often reasonable to assume that $Q\{Y(j) \geq Y(k)\} = 1$ for any pair of treatments $j, k \in \mathcal{D}$ with $j \geq k$. For example, in Manski and Pepper (1998), the treatment is years of schooling and the outcome is log(wage), and the restriction is that additional schooling leads to (weakly) higher wages. This restriction can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), \ldots, y(|\mathcal{D}| - 1), d(0), \ldots, d(|\mathcal{Z}| - 1)) : y(K_d) \geq \ldots \geq y(0)\}$. While Manski (1997) develops partial-identification results for unconditional treatment parameters in this context without restrictions on treatment-response types, our analysis additionally allows such restrictions to be incorporated and enables inference on parameters that condition on generalized principal strata. ■

Example 3.9 (*Harmless treatment*)**.** Consider a setting in which $Z \in \mathcal{Z} = \{0, \ldots, K_z\}$ is randomly assigned, so Assumption 2.1 holds, and there is a baseline treatment, i.e., the control, corresponding to $0 \in \mathcal{D} = \{0, \ldots, K_d\}$. In many applications, it is reasonable to assume that the remaining treatments are harmless relative to the control. For example, in Angrist et al. (2009), the outcome is academic performance, the control is no treatment, and the noncontrol treatments are (i) providing students with academic peer-advising service; (ii) financial incentives for good academic performance; and (iii) both (i) and (ii). It is therefore natural to assume that $Q\{Y(d) \geq Y(0)\} = 1$ for all $d \in \mathcal{D}$. This restriction can be formulated in terms of Assumption 2.2 with $\mathcal{R} = \{(y(0), \ldots, y(K_d), d(0), \ldots, d(K_z)) : y(d) \geq y(0) \text{ for all } d \in \mathcal{D}\}$. This restriction is an example of what Manski (1997) termed a semi-monotone ordering of outcomes. As discussed above in Example 3.8, our analysis differs from that of Manski (1997) by additionally allowing restrictions on treatment-response types to be incorporated and enabling inference on parameters that condition on generalized principal strata. ■

In many of the examples discussed above, Y may be an ordinal outcome. In such instances, average treatment effects may not be interpretable. Nonetheless, researchers may consider other parameters that fit our framework, such as: (i) $Q\{Y(j) > Y(k) \mid R \in \mathcal{R}'\}$, the conditional probability of benefit of treatment j versus treatment k; (ii) $Q\{Y(j) \geq Y(k) \mid R \in \mathcal{R}'\}$, the conditional probability of no harm of treatment j versus treatment k; or (iii) $Q\{Y(j) > Y(k) \mid R \in \mathcal{R}'\} - Q\{Y(k) > Y(j) \mid R \in \mathcal{R}'\}$, the conditional relative treatment effect of treatment j versus k. Each of these parameters can be written as (2) for appropriate choices of g. These parameters have been previously studied for the special case in which $\mathcal{R}' = \mathcal{R}$ and $|\mathcal{D}| = 2$ or 3 by Lu et al. (2018), Huang et al. (2019), and Gabriel et al. (2024).

It is also instructive to discuss examples of restrictions that *do not* fit our framework. Angrist and Imbens (1995), for instance, consider a generalization of the the monotonicity assumption in Example 3.7 in which the direction of the monotonicity is not known *a priori*, i.e., they consider the restriction that, for any pair of values for the instrument $j, k \in \mathcal{Z}$, either $Q\{D(j) \geq D(k)\} = 1$ or $Q\{D(j) \leq D(k)\} = 1$. Such a restriction is also analyzed in Vytlacil (2006) and in Heckman and Pinto (2018), where the latter paper terms it "ordered monotonicity." This restriction is equivalent to imposing $Q\{R \in \mathcal{R}_\pi\} = 1$ for some $\pi \in \Pi(|\mathcal{Z}|)$, where $\Pi(|\mathcal{Z}|)$ is the set of all permutations of $\{0, \ldots, |\mathcal{Z}| - 1\}$ and $\mathcal{R}_\pi = \{(y(0), \ldots, y(|\mathcal{D}| - 1), d(0), \ldots, d(|\mathcal{D}| - 1)) : d(\pi(|\mathcal{Z}| - 1)) \geq \ldots \geq d(\pi(0))\}$. A similar generalization of the assumption in Example 3.8 can also be considered. In particular, Machado et al. (2019) study such a generalization with $|\mathcal{D}| = |\mathcal{Z}| = 2$. These models, in which the permutation π is not known *a priori*, fall outside the class of models we consider.

4 Inference with Discrete Outcomes

In this section, we propose a test for conducting inference on the parameter of interest $\theta(Q)$. To this end, recall that P denotes the distribution of (Y, D, Z) and set $\mathcal{M} \equiv \mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$. In this section only, we suppose $|\mathcal{Y}| < \infty$. Doing so allows us to introduce our method succinctly. This assumption will be removed in Section 5, where we allow for a more general outcome through discretization. For $(y, d, z) \in \mathcal{M}$, define $P_{ydz} = P\{Y = y, D = d, Z = z\}$ and $P_z = P\{Z = z\}$. In what follows, we assume that $P \in \mathbf{P}$, where $\mathbf{P}$ is a "large" class of distributions that we will specify below. The class $\mathbf{P}$ can depend on the sample size n, but we suppress the dependence from our notation. We consider the problem of testing

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P} \setminus \mathbf{P}_0 \tag{10}$$

at level $\alpha \in (0, 1)$, where, for a pre-specified value of θ_0, the null hypothesis we consider is given by

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \theta_0 \in \Theta_0(P, \mathbf{Q})\} \ . \tag{11}$$

As we show below, a test of this null hypothesis can be used to construct confidence regions for $\theta(Q)$ through test inversion over the pre-specified value θ_0.

The key insight underlying the construction of our test is the following theorem, which provides a convenient reformulation of $\mathbf{P}_0$ in terms of existence of a non-negative solution to a (possibly under-determined) system of linear equations in which the "coefficients" are known. The statement of the theorem involves the parameter

$$\beta(P) \equiv ((P_{yd|z} : (y, d, z) \in \mathcal{M}), 1, 0)' \ , \tag{12}$$

where $P_{yd|z} \equiv P\{Y = y, D = d|Z = z\}$. Using this notation, we obtain the following result:

Theorem 4.1. *Suppose $|\mathcal{Y}| < \infty$. Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2, and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{Z = z\} > 0$. Then, for $\beta(P)$ defined in (12) and a matrix A defined in the beginning of Appendix A.1 that depends only on $\mathcal{R}$ in Assumption 2.2, θ_0 in (11), and the quantities g and $\mathcal{R}'$ in* *the definition of* $\theta(Q)$ *in* (2), *it follows*

$$\mathbf{P}_0 \subseteq \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} \ . \tag{13}$$

Furthermore, if $\mathbf{P}_0 \neq \emptyset$, *then* $\{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\} \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$, *where* $\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ *denotes the closure of* $\mathbf{P}_0$ *in* $\mathbf{P}$ *(understood as a subset of* $\mathbf{R}^{|\mathcal{M}|}$*).*

The key implication of Theorem 4.1 is that we may test the null hypothesis of interest by examining whether there exists a positive vector x satisfying $Ax = \beta(P)$. Because $\theta(Q)$ need not always be well defined when $\mathcal{R}' \neq \mathcal{R}$, (13) involves an inclusion rather than an equality. From a testing perspective, however, $\mathbf{P}_0$ and its closure are indistinguishable, in the sense that any level α test of whether P belongs to $\mathbf{P}_0$ necessarily has power no larger than α against any P in the closure $\text{cl}(\mathbf{P}_0)$ (Romano, 2004). Therefore, the second part of Theorem 4.1 may be interpreted as showing that testing whether $P \in \mathbf{P}_0$ is in fact equivalent to testing whether $\beta(P) = Ax$ for some $x \geq 0$. This conclusion holds under the condition that $\mathbf{P}_0$ is not empty. However, by definition, $\mathbf{P}_0$ is empty whenever there is no possible distribution of the data $P \in \mathbf{P}$ that is compatible with the null hypothesis. In other words, we may focus on testing whether $\beta(P) = Ax$ for some $x \geq 0$ *unless* the null hypothesis specifies restrictions that are incompatible with any P and, in this sense, may be interpreted as being logically inconsistent with each other.

Theorem 4.1 also has implications for deriving an analytical expression for the identified set of $\theta(Q)$. We discuss these implications in the next two remarks, and elaborate on them in the appendix.

Remark 4.1. When $Q\{R \in \mathcal{R}'\}$ is known or identified, in the sense that $\{Q\{R \in \mathcal{R}'\} : Q \in \mathbf{Q}_0(P, \mathbf{Q})\}$ is a singleton for all $P \in \mathbf{P}$ for which $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$, it is possible to use the characterization of $\mathbf{P}_0$ in Theorem 4.1 to derive closed-form expressions $L(P)$ and $U(P)$ such that $\Theta_0(P, \mathbf{Q}) = [L(P), U(P)]$ for all $P \in \mathbf{P}$ for which $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$ and $Q\{R \in \mathcal{R}'\} > 0$. As in Balke and Pearl (1997), the key idea is to express $L(P)$ and $U(P)$ as the values of linear programs and to use the duals of these programs to obtain expressions for $L(P)$ and $U(P)$ in terms of $P_{yd|z}$ through vertex enumeration. Computing $L(P)$ and $U(P)$ in this way rapidly becomes computationally prohibitive as the support of (Y, D, Z) becomes large. We describe this procedure in Appendix B and further develop a related procedure when $Q\{R \in \mathcal{R}'\}$ is not identified. Following the approach in Machado et al. (2019), it is possible to use such expressions for inference, but a virtue of the approach we pursue here is that it does not rely on knowledge of these expressions. ■

Remark 4.2. The identified set for the parameter of interest in general depends on the distribution of the full vector (Y, D, Z). This is the case even for parameters that only depend on the distribution of treatment response types R_t, $Q\{R_t = r_t\}$—parameters that, when point-identified, typically depend only on the distribution of (D, Z) (Heckman and Pinto, 2018; Navjeevan et al., 2023). In Appendix C, we illustrate this phenomenon by deriving the closed-form expression for the identified set in an example using the method described in Remark 4.1, and comparing it with the bounds using only the distribution of (D, Z). ■

Theorem 4.1 immediately permits us to test whether the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. To see this, consider $g \equiv 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$. For such values of these quantities,

$$\Theta_0(P, \mathbf{Q}) = \begin{cases} \{1\} & \text{if } \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset \\ \emptyset & \text{otherwise.} \end{cases} .$$

Therefore, $\mathbf{P}_0 = \{P \in \mathbf{P} : 1 \in \Theta_0(P, \mathbf{Q})\} = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}$. In this case, the equality in the last row of $Ax = \beta(P)$ in (13) is always true, so it can be omitted. In order to state the result formally, denote by $\tilde{A}$ and $\tilde{\beta}(P)$ all but the last rows of A and $\beta(P)$, respectively, in Theorem 4.1.

Corollary 4.1. *Suppose $|\mathcal{Y}| < \infty$. Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2, and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{Z = z\} > 0$. For $g \equiv 1$, $\theta_0 = 1$, and $\mathcal{R}' = \mathcal{R}$,*

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\} = \{P \in \mathbf{P} : \tilde{A}x = \tilde{\beta}(P) \text{ for some } x \geq 0\} . \qquad (14)$$

Because $\theta(Q)$ is always well defined when $\mathcal{R}' = \mathcal{R}$, (14) in Corollary 4.1 differs from its counterpart (13) in Theorem 4.1 in that it involves an equality instead of an inclusion.

Remark 4.3. Based on the characterization in Corollary 4.1, we show in Appendix D how a strategy like that described in Remark 4.1 can be used to derive a set of analytical inequalities that hold if and only if $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. We emphasize, however, that our inference methods below do not rely on these analytical inequalities. ■

By Theorem 4.1, we may test the null hypothesis of interest by testing whether P is such that

$$\beta(P) = Ax \text{ for some } x \geq 0. \qquad (15)$$

In recent work, Fang et al. (2023) derived a test for whether P satisfies the restriction in

(15) for a general class of problems; see also Bai et al. (2022) for other approaches. In what follows, we build on Fang et al. (2023) by employing the structure of our problem to sharpen rate conditions and establish the validity of their test under weaker requirements. Following Corollary 4.1, it is straightforward to modify the test described below to test the null hypothesis that the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$.

In order to describe the test statistic, we require some additional notation. We will assume the availability of an i.i.d. sample $\{V_i\}_{i=1}^n$ with $V_i = (Y_i, D_i, Z_i)$, $V_i \sim P$, and let $\hat{P}_n$ denote the empirical distribution corresponding to the sample $\{V_i\}_{i=1}^n$. We further set $\hat{\beta}_n \equiv \beta(\hat{P}_n)$ to denote the plug-in estimator for the parameter $\beta(P)$ and let $\hat{\Omega} \equiv \Omega(\hat{P}_n)$, where $\Omega(P)$ is a diagonal weighting matrix whose (i,i)-entry equals the asymptotic standard deviation of the i^{th} row of $\sqrt{n}(\hat{\beta}_n - \beta(P))$. Specifically, the first $|\mathcal{M}|$ diagonal entries of $\Omega(P)$ equal

$$\left(\frac{P_{yd|z}(1 - P_{yd|z})}{P_z}\right)^{1/2}$$

for some $(y, d, z) \in \mathcal{M}$, and the final two diagonal entries of $\Omega(P)$ are equal to zero. Finally, we let $A^\dagger$ denote the Moore-Penrose pseudoinverse of A.

Given the introduced notation, we define the set $\hat{\mathcal{V}}_n \equiv \{s : A^\dagger s \leq 0, \|\hat{\Omega}_n (AA')^\dagger s\|_1 \leq 1\}$ and set

$$T_n = \sup_{s \in \hat{\mathcal{V}}_n} \sqrt{n} \langle A^\dagger s, A^\dagger \hat{\beta}_n \rangle \tag{16}$$

as our test-statistic – here, $\|a\|_1 = \sum_{i=1}^d |a_i|$ for any vector $a = (a_1, \ldots, a_d) \in \mathbf{R}^d$ and $\langle a, b \rangle = a'b$. From results in Fang and Santos (2019), T_n represents a sample analogue to an equivalent formulation of the restriction in (15) obtained from Farkas' lemma. In particular, the population analogue to T_n equals zero whenever (15) indeed holds and, as a result, under the null hypothesis T_n should not be "too large." We highlight that the test statistic can be computed through linear programming. As a result, T_n can be reliably and quickly computed even in high-dimensional applications.

Unfortunately, the asymptotic distribution of T_n is not pivotal. In particular, its asymptotic distribution depends on the "directions" $s \in \hat{\mathcal{V}}_n$ at which the population analogue $\langle A^\dagger s, A^\dagger \beta(P) \rangle$ is "close" to zero. In order to obtain a valid critical value, we therefore rely on the nonparametric bootstrap and a construction that asymptotically excludes directions s that do not play a role in the distribution of T_n; i.e., directions s for which $\langle A^\dagger s, A^\dagger \beta(P) \rangle$ is "far" from zero. Specifically, for the latter purpose we introduce

$$\hat{\beta}_n^r \in \operatorname*{argmin}_b \sup_{s \in \hat{\mathcal{V}}_n} |\langle A^\dagger s, A^\dagger (\hat{\beta}_n - b) \rangle|$$

$$\text{s.t. } Ax = b \text{ for some } x \geq 0 \text{ and } b = (b_u', 1, 0)' \text{ for } b_u \in \mathbf{R}^{|\mathcal{M}|},$$

which represents an estimator for $\beta(P)$ that is restricted to satisfy the null hypothesis. Letting $\{V_i^*\}_{i=1}^n$ denote a bootstrap sample, (i.e., a sample drawn i.i.d. from $\hat{P}_n$), and $\hat{\beta}_n^*$ the bootstrap analogue to $\hat{\beta}_n$, we set

$$T_n^* \equiv \sup_{s \in \hat{\mathcal{V}}_n} \left\{ \sqrt{n} \langle A^\dagger s, A^\dagger (\hat{\beta}_n^* - \hat{\beta}_n) \rangle + \lambda_n \sqrt{n} \langle A^\dagger s, A^\dagger \hat{\beta}_n^r \rangle \right\} \tag{17}$$

as our bootstrap statistic. Here, $\lambda_n \in [0,1]$ represents a penalty satisfying $\lambda_n = o(1)$, and whose role is to ensure that directions s that do not impact the finite-sample distribution of our test statistic do not impact the bootstrap statistic either. We again highlight that, just like T_n, the bootstrap statistic T_n^* can also be computed through linear programming.

The critical value for a level α test is then simply the $1-\alpha$ quantile of the distribution of T_n^* conditionally on the sample $\{V_i\}_{i=1}^n$. Formally, the critical value for our test is given by

$$\hat{c}_n(1-\alpha) \equiv \inf \left\{ u \in \mathbf{R} : \hat{P}_n \{ T_n^* \leq u \} \geq 1 - \alpha \right\}. \tag{18}$$

As usual, the critical value $\hat{c}_n(1-\alpha)$ can be computed through simulation by drawing multiple samples $\{V_i^*\}_{i=1}^n$ i.i.d. from $\hat{P}_n$ and computing the $1-\alpha$ quantile of the corresponding sample of bootstrap statistics. Given the critical value, we next formally introduce our test, which rejects whenever T_n exceeds $\hat{c}_n(1-\alpha)$:

$$\phi_n(\theta_0) \equiv I\{T_n > \hat{c}_n(1-\alpha)\} \ . \tag{19}$$

We establish the asymptotic validity of our test under the following main assumption.

Assumption 4.1. (i) $V_i = (Y_i, D_i, Z_i), i = 1, \ldots, n$ is an i.i.d. sample with $V_i \sim P \in \mathbf{P}$; (ii) $(\hat{\beta}_n - \beta(P)) \in \text{range}(A)$ with probability tending to one uniformly in $P \in \mathbf{P}$; (iii) For some constants $0 < c_1 < c_2$,

$$c_1 \leq \inf_{P \in \mathbf{P}} \min_{(y,d,z) \in \mathcal{M}} |\mathcal{M}| P_{ydz} \leq \sup_{P \in \mathbf{P}} \max_{(y,d,z) \in \mathcal{M}} |\mathcal{M}| P_{ydz} \leq c_2.$$

Assumption 4.1(i) simply formalizes our requirement that the sample be independent and identically distributed. In Assumption 4.1(ii) we impose the additional condition that $(\hat{\beta}_n - \beta(P))$ belong to the range of the matrix A with probability tending to one. This requirement is satisfied in our examples, where $(\hat{\beta}_n - \beta(P))$ belongs to the range of A whenever $\hat{P}_z > 0$ for all $z \in \mathcal{Z}$ – i.e., whenever $\hat{\beta}_n$ is well defined. We note that Assumption

4.1(ii) is not imposed in Fang et al. (2023), but we deploy it here due to its wide applicability and its importance in obtaining simple primitive conditions for establishing the validity of our test. In applications in which Assumption 4.1(ii) is violated, it is still possible to conduct inference by using the results in Fang et al. (2023). Finally, Assumption 4.1(iii) represents our main regularity condition on the set $\mathbf{P}$. Assumption 4.1(iii) requires that the probability of each support point (y, d, z) be proportional to each other. We note that this restriction implies that no conditional probability is dominant, in the sense that $P_{yd|z}$ is bounded away from one uniformly in the (y, d, z) and $P \in \mathbf{P}$.

We next establish the validity of our test under Assumption 4.1 and an additional regularity condition that we discuss after the theorem.

Theorem 4.2. *Suppose Assumption 4.1 holds, $\alpha \in (0, 0.5)$, $\log^3(n)|\mathcal{M}|/n = o(1)$, and $0 \leq \lambda_n \leq 1$ satisfies $\lambda_n\sqrt{\log(|\mathcal{M}|)} = o(1)$. Further suppose Assumption A.1 in the appendix holds. Then, the test in (19) satisfies*

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] \leq \alpha.$$

Theorem 4.2 establishes the validity of our test under weak conditions on the number of support points for (Y, D, Z). The main rate requirement imposed by Theorem 4.2 is that $|\mathcal{M}|/n$ tend to zero (up to logs). We view this rate condition as close to minimal, in the sense that it is necessary for the empirical distribution $\hat{P}_n$ to be consistent for P. Our rate restriction is in addition a significant improvement over Fang et al. (2023), whose best attainable rate is that $|\mathcal{M}|^2/n$ tend to zero (up to logs). We improve on their results by employing the specific structure of our problem to derive better coupling rates for our test and bootstrap statistics. In particular, by relying on results in Massart (1989) we are able to couple our test statistic to a Gaussian counterpart at a rate r_n given by

$$r_n \equiv \frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}.$$

The bootstrap statistic is in turn shown to be coupled to a Gaussian distribution at a rate of $r_n^{1/2}$; see Lemma E.3 in the Appendix for additional details. We conjecture that the bootstrap statistic can in fact be coupled at a rate r_n as well. However, we do not pursue this refinement because it does not impact the conditions on how $|\mathcal{M}|$ relates to n, which is our primary concern.

Showing weak convergence of a test statistic and its bootstrap counterpart to a common distribution is in general not sufficient for establishing consistency of the corresponding critical values. In high-dimensional settings, consistency of the critical values requires a condition termed *anti-concentration* by Chernozhukov et al. (2015). Intuitively, anti-concentration ensures that the c.d.f. of the bootstrap statistic is suitably well behaved at the quantile of interest. Assumption A.1, stated in the Appendix for ease of exposition, is a sufficient condition for anti-concentration to hold in our setting. We note that Assumption A.1 is automatically satisfied in an asymptotic framework in which $|\mathcal{M}|$ is fixed with the sample size. Alternatively, Assumption A.1 can be dispensed with by modifying our critical value to include a "correction factor" introduced by Andrews and Shi (2013). We opt to introduce Assumption A.1 instead, however, to avoid introducing an additional tuning parameter.

Finally, we emphasize that while we have focused our discussion on hypothesis testing, our results readily deliver confidence regions as well. In particular, through test inversion it is straightforward to show that

$$C_n = \{\theta_0 \in \mathbf{R} : \phi_n(\theta_0) = 0\} \tag{20}$$

is a valid confidence region, in the sense that it covers the parameter of interest with asymptotic probability of $1 - \alpha$ uniformly in $P \in \mathbf{P}$.

Remark 4.4. Assumption 2.2 imposes that response types take on certain values with probability zero. Our framework can easily be modified, however, to accommodate the restriction that response types take on certain values with at most (or at least) some pre-specified probability. This modification is useful for conducting sensitivity analysis, as in Masten and Poirier (2020) and Kline and Santos (2013). For instance, in Example 3.3, one may relax restriction (4) so that it instead specifies that the left-hand side is at most some pre-specified amount $\epsilon > 0$, and explore how inferences on $\theta(Q)$ change as one varies ϵ. In this way, we can, for example, generalize the analysis of Noack (2021) to multi-valued instruments and treatments. ■

Remark 4.5. In applications with closed-form expressions of the identified set, $\Theta_0(P, \mathbf{Q}) = [L(P), U(P)]$, the plug-in estimator $[L(\hat{P}_n), U(\hat{P}_n)]$ provides a natural approach for estimating the identified set. However, the plug-in bootstrap in general does not provide correct coverage. Formally, denote by $\ell(\alpha/2, P)$ the $\alpha/2$ quantile of the distribution of $L(\hat{P}_n)$ under P and by $u(1-\alpha/2, P)$ the $1-\alpha/2$ quantile of the distribution of $U(\hat{P}_n)$ under P. Consider the confidence region given by the bootstrap analogue $[\ell(\alpha/2, \hat{P}_n), u(1-\alpha/2, \hat{P}_n)]$. Because the functionals $L(P)$ and $U(P)$ generally involve the maximum and minimum over functionals of P, they are only directionally differentiable and not fully differentiable when the maximum and minimum are attained at multiple functionals. Hence, it follows from results in Fang and Santos (2019) that the bootstrap quantiles $\ell(\alpha/2, \hat{P}_n)$ and $u(1-\alpha/2, \hat{P}_n)$ are not consistent for the population quantiles $\ell(\alpha/2, P)$ and $u(1-\alpha/2, P)$ that justify the validity of the proposed confidence region. ■

5 Inference with General-valued Outcomes

Section 4 proposed an inference framework with the support $\mathcal{Y}$ of Y restricted to be finite. We now extend that framework by relaxing the restriction that $\mathcal{Y}$ be finite, in particular allowing for continuous outcomes. The treatment D and instrument Z remain discrete with supports $\mathcal{D}$ and $\mathcal{Z}$ satisfying $2 \leq |\mathcal{D}|, |\mathcal{Z}| < \infty$.

Our approach essentially partitions the outcome space $\mathcal{Y}$ into a finite number of subsets $\{B_{\ell,L}\}_{\ell=1}^{L}$ and reduces the problem into a discrete one. The analysis mimics the discrete case studied in Theorem 4.1, but replaces the parameter $\beta(P)$ by a vector $\beta_L(P)$ that is given by

$$\beta_L(P) \equiv ((P\{Y \in B_{\ell,L}, D = d|Z = z\} : (\ell, d, z) \in \mathcal{L} \times \mathcal{D} \times \mathcal{Z}), 1, 0, 0)' , \tag{21}$$

where $\mathcal{L} \equiv \{1, \ldots, L\}$. In what follows, it is helpful to note that because $R_o \in \mathcal{Y}^{|\mathcal{D}|}$, any partition $\{B_{\ell,L}\}_{\ell=1}^{L}$ induces a partition for the support of R_o as well. Specifically, writing any $c_o \in \mathcal{L}^{|\mathcal{D}|}$ as $c_o \equiv (c_o(0), \ldots, c_o(|\mathcal{D}|-1))$, and setting $B_{c_o} \equiv B_{c_o(0),L} \times \ldots \times B_{c_o(|\mathcal{D}|-1)}$ we have that $\{B_{c_o}\}_{c_o \in \mathcal{L}^{|\mathcal{D}|}}$ forms a partition for $\mathcal{Y}^{|\mathcal{D}|}$. Our next assumption uses this notation in introducing the main regularity conditions we require:

Assumption 5.1. (i) $\mathcal{Y}$ is compact; (ii) g is continuous; (iii) $\{B_{\ell,L}\}_{\ell=1}^{L}$ is a partition of $\mathcal{Y}$ satisfying $\max_{1\leq \ell \leq L} \operatorname{diam}\{B_{\ell,L}\} = o(1)$ as $L \to \infty$; (iv) $\mathcal{R}'_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|\mathcal{D}|} : (r_o, r_t) \in \mathcal{R}'\}$ is closed and $B_{c_o} \cap \mathcal{R}'_o(r_t)$ is connected for all $c_o \in \mathcal{L}^{|\mathcal{D}|}$ and $r_t \in \mathcal{D}^{|\mathcal{Z}|}$.

Assumption 5.1(i) imposes that $\mathcal{Y}$ be closed and bounded, which ensures that the partition $\{B_{\ell,L}\}_{\ell=1}^{L}$ can be chosen so that each set $B_{\ell,L}$ is bounded. In Assumption 5.1(ii), we further restrict attention to the case in which g is continuous. As we discuss in Remark 5.1 below, some forms of discontinuity can be accommodated, though they can require specific choices of partitions $\{B_{\ell,L}\}_{\ell=1}^{L}$ that satisfy restrictions beyond those imposed by Assumption 5.1. Finally, Assumptions 5.1(iii)-(iv) impose the main requirements on the partition $\{B_{\ell,L}\}_{\ell=1}^{L}$, which include that it become finer as L becomes large. In applications in which $\mathcal{R}$ and $\mathcal{R}'$ leave the outcome response type R_o unrestricted, $\mathcal{R}'_o(r_t) = \mathcal{Y}^{|\mathcal{D}|}$ and Assumption 5.1(iv) is satisfied whenever the sets $\{B_{\ell,L}\}_{\ell=1}^{L}$ are chosen to be convex.

The next result may be seen as a generalization of Theorem 4.1.

Theorem 5.1. *Let* $\mathbf{Q}$ *be the set of all distributions of* (R, Z) *satisfying Assumptions 2.1 and 2.2 and* $\mathbf{P}$ *be the set of all distributions of* (Y, D, Z) *satisfying* $P\{Z = z\} > 0$ *for every* $z \in \mathcal{Z}$. *Further suppose* g *in the definition of* $\theta(Q)$ *in (2) is bounded. Then, for* $\beta_L(P)$ *defined in (21) and a matrix* A_L *defined in the beginning of Appendix A.3 that depends only on* $\mathcal{R}$ *in Assumption 2.2,* θ_0 *in (11), and* $\mathcal{R}'$ *in the definition of* $\theta(Q)$ *in (2), it follows that*

$$\mathbf{P}_0 \subseteq \{P \in \mathbf{P} : A_L x = \beta_L(P) \text{ for some } x \geq 0\} \equiv \mathbf{P}_{0,L} \ . \tag{22}$$

If in addition Assumption 5.1 holds and $\mathbf{P}_0$ *is not empty, then*

$$\limsup_{L \to \infty} \mathbf{P}_{0,L} \equiv \bigcap_{M=1}^{\infty} \bigcup_{L \geq M} \mathbf{P}_{0,L} \subseteq \mathrm{cl}_{\mathbf{P}}(\mathbf{P}_0) \ ,$$

where $\mathrm{cl}_{\mathbf{P}}(\mathbf{P}_0)$ *denotes the closure of* $\mathbf{P}_0$ *in* $\mathbf{P}$ *under weak convergence.*

The first part of Theorem 5.1 shows that $\mathbf{P}_0 \subseteq \mathbf{P}_{0,L}$ for $\mathbf{P}_{0,L}$ as defined in (22). Therefore, any test of the null hypothesis that $P \in \mathbf{P}_{0,L}$ is also a valid test for the null hypothesis that $P \in \mathbf{P}_0$. Importantly, this conclusion does not depend on any of the regularity conditions imposed in Assumption 5.1. The second part of Theorem 5.1 shows, under Assumption 5.1, that $\limsup_{L \to \infty} \mathbf{P}_{0,L}$ is included in the closure of $\mathbf{P}_0$. The second part of Theorem 5.1 therefore shows that, as our discretization becomes finer, we are able to detect whether P belongs to $\mathbf{P}_0$ or not (up to closure).

The main implication of Theorem 5.1 is that, with a continuous outcome, we still can test the null hypothesis of interest by testing whether P is such that $A_L x = \beta_L(P)$ for some $x \geq 0$. The latter null hypothesis can be tested by using an analogous approach to that employed in (19) for the finite support case. The asymptotic validity of the resulting test holds under similar assumptions to those required in Theorem 4.2, but applied with $\mathcal{M}_L \equiv \mathcal{L} \times \mathcal{D} \times \mathcal{Z}$ instead of $\mathcal{M}$. This requires, among other assumptions, that $|\mathcal{M}_L|/n$ tend to zero (up to logs). In particular, if $\mathcal{D}$ and $\mathcal{Z}$ are fixed, then the condition that $|\mathcal{M}_L|/n$ tend to zero is equivalent to requiring that L/n tending to zero. The partition $\{B_{\ell,L}\}_{\ell=1}^{L}$ is therefore allowed to quickly become finer with the sample size which, in light of Theorem 5.1, can be desirable for power considerations.

Remark 5.1. An important example of a parameter g that corresponds to a discontinuous g, and hence violates Assumption 5.1(i), is the probability of an event conditional on generalized principal strata. However, in this case, g is piecewise constant and the conclusion of the second part of Theorem 5.1 continues to hold if we redefine A_L suitably. To see why, note by an inspection of the proof of Theorem 5.1 that a solution to the linear system in (22) is a vector of probabilities the cells determined by the intersections of $B_{c_o} \times \{r_t\}$ and $\mathcal{R}'$. If $\mathcal{Y}^{|\mathcal{D}|} = \bigcup_{1 \leq j \leq K} \mathcal{S}_j$ and g is constant on $\mathcal{S}_j$ for each j, then we redefine A_L so that the solution to the linear system is instead a vector of probabilities on each $B_{c_o} \times \{r_t\}$ and $\mathcal{R}'$ with $\mathcal{S}_j$ for $1 \leq j \leq K$. ■

By arguing as in Section 4, Theorem 5.1 immediately permits us to test whether the model is correctly specified in the sense that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. Recall that by setting $g \equiv 1$, $\theta_0 = 1$ and $\mathcal{R}' = \mathcal{R}$, we have that $\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\}$. In this case, the equalities in the last two rows of $A_L x = \beta_L(P)$ in (22) are always true, so they can be omitted. In order to state the result, denote by $\tilde{A}_L$ and $\tilde{\beta}_L(P)$ all but the last two rows of A_L and $\beta_L(P)$, respectively, in Theorem 5.1.

Corollary 5.1. *Let $\mathbf{Q}$ be the set of all distributions of (R, Z) satisfying Assumptions 2.1 and 2.2 and $\mathbf{P}$ be the set of all distributions of (Y, D, Z) satisfying $P\{Z = z\} > 0$ for every $z \in \mathcal{Z}$. For $g \equiv 1$, $\theta_0 = 1$ and $\mathcal{R}' = \mathcal{R}$,*

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset\} \subseteq \{P \in \mathbf{P} : \tilde{A}_L x = \tilde{\beta}_L(P)\} \equiv \tilde{\mathbf{P}}_{0,L} \ .$$

If in addition Assumption 5.1 holds and $\mathbf{P}_0$ is not empty, then $\limsup_{L \to \infty} \tilde{\mathbf{P}}_{0,L} \equiv \bigcap_{M=1}^{\infty} \bigcup_{L \geq M} \tilde{\mathbf{P}}_{0,L}$ satisfies $\limsup_{L \to \infty} \tilde{\mathbf{P}}_{0,L} \subseteq \mathrm{cl}_{\mathbf{P}}(\mathbf{P}_0)$ where $\mathrm{cl}_{\mathbf{P}}(\mathbf{P}_0)$ denotes the closure of $\mathbf{P}_0$ in $\mathbf{P}$ under weak convergence.

Remark 5.2. As in Remark 4.1, when $Q\{R \in \mathcal{R}'\}$ is known or identified for each $L \geq 1$, it is possible to use the characterization of $\mathbf{P}_0$ in Theorem 5.1 to derive *closed-form expressions* $L_L(P)$ and $U_L(P)$ such that $\Theta_0(P, \mathbf{Q}) \subseteq [L_L(P), U_L(P)]$ for all $P \in \mathbf{P}$ for which $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. The inclusion $\Theta_0(P, \mathbf{Q}) \subseteq [L_L(P), U_L(P)]$ is generally strict for each fixed L because of the loss of information from discretization. Therefore, as opposed to settings in which the support of Y is discrete, for a fixed L we typically obtain closed-form expressions for an outer set of the identified set instead of the identified set itself. The second part of Theorem 5.1 implies, however, that under appropriate assumptions $\limsup_{L \to \infty} [L_L(P), U_L(P)] = \Theta_0(P, \mathbf{Q})$. Similarly, as in Remark 4.3, for each $L \geq 1$, it is possible to use Corollary 5.1 to obtain a set of analytical inequalities $a_L(P) \leq 0$, such that $P \in \mathbf{P}$ satisfies $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$ if and only if $a_L(P) \leq 0$ for every $L \geq 1$. ■

6 Simulations

In this section we present simulation results that illustrate the finite-sample performance of our proposed inference methods. Throughout this section, $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$. We set $\mathcal{Y}$ to be discrete in Section 6.1 and an interval in Section 6.2. The parameters of interest are

$$\theta_{\text{ATE}}(Q) = E_Q[Y(2) - Y(1) \mid R \in \mathcal{R}'] \quad \text{and} \quad \theta_{\text{Prob}}(Q) = Q\{Y(2) \geq Y(1) \mid R \in \mathcal{R}'\} ,$$

where $\mathcal{R}' = \{(y(0), y(1), y(2), d(0), d(1), d(2)) : (d(0), d(1), d(2)) = (0, 1, 2)\}$. We consider the following two specifications of $\mathcal{R}$, corresponding to two different models:

One-sided noncompliance: $\mathcal{R}$ is as specified in Example 3.1 and denoted by $\mathcal{R}_{\text{1s}}$, with the corresponding model $\mathbf{Q}_{\text{1s}} = \{Q : Q$ satisfies Assumptions 2.1 and 2.2 with $\mathcal{R} = \mathcal{R}_{\text{1s}}\}$;

Encouragement design: $\mathcal{R}$ is as specified in Example 3.3 and denoted by $\mathcal{R}_{\text{enc}}$, with the corresponding model $\mathbf{Q}_{\text{enc}} = \{Q : Q$ satisfies Assumptions 2.1 and 2.2 with $\mathcal{R} = \mathcal{R}_{\text{enc}}\}$.

6.1 Discrete Outcomes

We first present simulation results for a discrete outcome with $\mathcal{Y} = \{-1, 0, 1\}$. From each of $\mathbf{Q}_{\text{1s}}$ and $\mathbf{Q}_{\text{enc}}$, we pick a Q distribution and fix them. The two selected Q's are described in Appendices F.1 and F.2, respectively. For each Q, we set $Q\{Z = z\} = 1/3$ for each $z \in \mathcal{Z}$ and define $P = QT^{-1}$. The simulation goes as follows. For each Q, we repeatedly draw an i.i.d. sample $\{V_i\}_{i=1}^n$ of size $n = 3000$ or 12000 from P, where $V_i = (Y_i, D_i, Z_i)$. For each sample, we obtain the confidence interval for θ_{ATE} and θ_{Prob} by inverting the test of the feasibility of the linear system in (13) at the 5% level.

The results are reported in Tables 1 and 2 for $\mathbf{Q}_{\text{1s}}$ and $\mathbf{Q}_{\text{enc}}$ respectively. For each parameter and each model, we report the length and the coverage rate of the confidence interval in the last two columns, averaged over 3000 replications. We also report the length of the identified set $|\Theta_0(P, \mathbf{Q}_{\text{1s}})|$ or $|\Theta_0(P, \mathbf{Q}_{\text{enc}})|$. We note that the confidence interval covers the parameter with probability one in all cases. This phenomenon is to be expected because in each case the true value of the parameter lies strictly in the interior of the identified set. For example, $\Theta_0(P, \mathbf{Q}_{1s}) = [0.1482, 1.0295]$ and $\theta_{\text{ATE}}(Q) = 0.4919$. In this example, for $n = 3000$, the coverage probabilities of the lower and upper endpoints of the identified set are 0.947 and 0.956, respectively.

6.2 General-valued Outcomes

We now present simulation results for a continuous outcome. In this subsection, $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$ and $\mathcal{Y} = [-5, 5]$. We focus on $\mathcal{R} = \mathcal{R}_{\text{1s}}$ and $\theta_{\text{ATE}}(Q) = E_Q[Y(2) - Y(1) \mid R \in \mathcal{R}']$. As in Section 6.1, $Q\{Z = z\} = 1/3$ for each $z \in \mathcal{Z}$. A distribution $Q \in \mathbf{Q}_{\text{1s}}$ is generated as

	$\lvert\Theta_0(P, \mathbf{Q}_{1s})\rvert$	n	95% CI length	Coverage
θ_{ATE}	0.8813	3000	1.034	1.000
		12000	0.961	1.000
θ_{Prob}	0.3805	3000	0.413	1.000
		12000	0.397	1.000

Table 1: Length of the identified set $\Theta_0(P, \mathbf{Q}_{1s})$, average length and average coverage rate of the 95% confidence intervals for θ_{ATE} and θ_{Prob} using our inference method in Section 4 under the $Q \in \mathbf{Q}_{1s}$ distribution specified in Appendix F.1.

	$\lvert\Theta_0(P, \mathbf{Q}_{\text{enc}})\rvert$	n	95% CI length	Coverage
θ_{ATE}	2.8355	3000	3.340	1.000
		12000	3.085	1.000
θ_{Prob}	0.8935	3000	0.984	1.000
		12000	0.969	1.000

Table 2: Length of the identified set $\Theta_0(P, \mathbf{Q}_{1s})$, average length and average coverage rate of the 95% confidence intervals for θ_{ATE} and θ_{Prob} using our inference method in Section 4 under the $Q \in \mathbf{Q}_{1s}$ distribution specified in Appendix F.2.

follows. With a slight abuse of notation, let $\tilde{Q}$ on $\{-1, 0, 1\}^3 \times \mathcal{D}^3$ be the discrete distribution in Appendix F.1. We then define Q as a mixture of truncated normal distributions where the mixture weights are given by $\tilde{Q}$. In particular, let $\Phi_{\mathcal{Y}}(\mu(d))$ denote the truncation of normal distribution $N(\mu(d), 1)$ on $\mathcal{Y}$. Then,

$$Q = \bigoplus_{\substack{(\mu(0),\mu(1),\mu(2),d(0),d(1),d(2))\\ \in\{-1,0,1\}^3\times\mathcal{D}^3}} \left(\tilde{Q}(\mu(0), \mu(1), \mu(2), d(0), d(1), d(2)) \times \bigotimes_{d\in\mathcal{D}} \Phi_{\mathcal{Y}}((\mu(d)) \bigotimes_{z\in\mathcal{Z}} I\{d(z)\}\right). \tag{23}$$

In words, for each mixture component, the value of $D(z)$ is fixed at $d(z)$ for $z \in \mathcal{Z}$ and independently across $d \in \mathcal{D}$, $Y(d)$ has a truncated normal distribution. Finally let $P = QT^{-1}$. As in Section 6.1, we repeatedly draw an i.i.d. sample $\{V_i\}_{i=1}^n$ of size $n = 3000$ or 12000 from P, where $V_i = (Y_i, D_i, Z_i)$, and for each sample obtain confidence intervals by inverting the test of the feasibility of the linear system in (22) at the 5% level. We consider the following three partitions of $\mathcal{Y}$:

`bin1`: $\mathcal{Y} = [-5, 0) \cup [0, 5]$;

`bin2`: $\mathcal{Y} = [-5, -3) \cup [-3, -1) \cup \cdots \cup [3, 5]$;

`bin3`: $\mathcal{Y} = [-5, -4) \cup [-4, -3) \cup \cdots \cup [4, 5]$.

The results are reported in Table 3. For each level of discretization, we report the length and the coverage rate of the confidence intervals in the last two columns, averaged over 3000

Partition	dim(A)	$\lvert\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})\rvert$	n	95% CI length	Coverage
`bin1` $(L=2)$	$(10+3)\times(32+2)$	9.988	3000	9.995	1.000
			12000	9.994	1.000
`bin2` $(L=5)$	$(25+3)\times(500+2)$	5.772	3000	6.099	1.000
			12000	5.921	1.000
`bin3` $(L=10)$	$(50+3)\times(4000+2)$	3.656	3000	4.077	1.000
			12000	3.851	1.000

Table 3: Dimensionality of the test, length of $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$, and average length and coverage rates of the 95% confidence intervals using our inference method in Section 5, under given partitions with $L = 2, 5$ or 10.

replications. To illustrate the computational costs, we report in the second column of Table 3 the number of rows and columns of the A matrix in (22) for each level of discretization, which correspond to the number of constraints and the number of decision variables in each of the linear systems being tested. Next, recall Theorem 5.1 shows that for each level of discretization L, $\mathbf{P}_{0,L}$ contains $\mathbf{P}_0$. Let $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$ be the set of $\theta_0 \in \mathbf{R}$ such that $\mathbf{P}_{0,L}$ is nonempty. Theorem 5.1 implies that $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$ is a superset of the identified set $\Theta_0(P, \mathbf{Q}_{1\text{s}})$. In the third column of Table 3, we display the length of $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$ for each L. A comparison of the length of our confidence intervals to the length of $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$ reveals how much of the length of the confidence interval is attributed to sampling uncertainty.

We note the following findings from Table 3. First, as in Tables 1 and 2, the confidence intervals cover the parameter with probability one for each level of discretization we consider. This phenomenon is again to be expected because in each case the true value of the parameter lies strictly in the interior of $\Theta_{0,L}(P, \mathbf{Q}_{1\text{s}})$; coverage probabilities of the lower and upper endpoints of $\Theta_{0,L}(P, \mathbf{Q}_{1s})$ are approximately equal to the nominal level. Next, as the number of intervals in the partition increases, the dimensionality of the problem grows exponentially. Specifically, the number of constraints and decision variables scale at the rates $O(|\mathcal{L}||\mathcal{D}||\mathcal{Z}|)$ and $O(|\mathcal{L}|^{|\mathcal{D}|}|\mathcal{R}_t|)$, respectively. For a fixed sample size, a finer partition yields a shorter confidence interval, although we note it comes at a higher computational cost.

A Proof of Main Results

To reduce notational clutter for the appendix, we suppress the dependence of $\hat{P}_n$ on n and write $\hat{P}$ instead.

A.1 Proof of Theorem 4.1

We begin by formally defining the matrix A. Recall that $\mathcal{M} \equiv \mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$ and let $\mathcal{N} \equiv \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$. We now define a matrix A with $|\mathcal{M}| + 2$ rows and $|\mathcal{N}|$ columns. In order to describe the matrix A, index the first $|\mathcal{M}|$ rows of A by $(y, d, z) \in \mathcal{M}$ and the columns of A by $r = ((y(d) : d \in \mathcal{D}), (d(z) : z \in \mathcal{Z})) \in \mathcal{N}$. The $(y, d, z) \times r$ element of A is given by $I\{y(d) = y, d(z) = d\}$, and the $(|\mathcal{M}| + 1) \times r$ element of A is given by $I\{r \in \mathcal{R}\}$. Finally, the $(|\mathcal{M}| + 2) \times r$ element of A is given by $(g(r) - \theta_0)I\{r \in \mathcal{R}'\}$.

Given the introduced definitions, we next prove Theorem 4.1.

Proof of Theorem 4.1. Suppose $P \in \mathbf{P}_0$. Then, $\theta_0 \in \Theta_0(P, \mathbf{Q})$, so there is $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ such that $\theta_0 = \theta(Q)$ and $Q\{R \in \mathcal{R}'\} > 0$. Recall $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ if and only if $Q \in \mathbf{Q}$ and $P = QT^{-1}$. Since $Q \in \mathbf{Q}$, it satisfies both Assumptions 2.1 and 2.2. Because Q satisfies Assumption 2.1 and $P = QT^{-1}$, we have

$$P_{yd|z} \equiv P\{Y = y, D = d|Z = z\} = Q\{Y(d) = y, D(z) = d|Z = z\} = Q\{Y(d) = y, D(z) = d\} . \tag{24}$$

Let q be the column vector with $|\mathcal{N}|$ elements indexed by r, where the r element is $q(r) = Q\{R = r\}$. It follows from (24) and the definition of A that the first $|\mathcal{M}|$ rows of Aq equal the first $|\mathcal{M}|$ rows of $\beta(P)$. Next, note that because $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ implies Q satisfies Assumption 2.2, it follows that

$$\sum_{r \in \mathcal{R}} q(r) = 1 ,$$

and therefore the $|\mathcal{M}| + 1$ row of Aq equals the $|\mathcal{M}| + 1$ row of $\beta(P)$ by definition of A and $\beta(P)$. Finally, since $\theta_0 = \theta(Q)$ by hypothesis, the definition of the conditional expectation yields that

$$\theta_0 = \frac{E_Q[g(R)I\{R \in \mathcal{R}'\}]}{Q\{R \in \mathcal{R}'\}} , \tag{25}$$

where recall Q satisfies $Q\{R \in \mathcal{R}'\} > 0$. Rearranging, we obtain $E_Q[(g(R) - \theta_0)I\{R \in \mathcal{R}'\}] = 0$, so that

$$\sum_{r \in \mathcal{R}'} (g(r) - \theta_0)q(r) = 0 ,$$

and it follows from the definition of A that the $|\mathcal{M}|+2$ row of Aq equals the $|\mathcal{M}|+2$ row of $\beta(P)$ holds. Hence, we have shown that $\mathbf{P}_0 \subseteq \tilde{\mathbf{P}}_0$ for $\tilde{\mathbf{P}}_0 \equiv \{P \in \mathbf{P} : Ax = \beta(P) \text{ for some } x \geq 0\}$.

To conclude, it only remains to show that $\tilde{\mathbf{P}}_0 \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0) =$ whenever $\mathbf{P}_0 \neq \emptyset$. To this end, fix any $\tilde{P} \in \tilde{\mathbf{P}}_0$. Since $Aq = \beta(\tilde{P})$ for some $q \geq 0$, the definition of A and $\beta(\tilde{P})$ imply there exists a $\tilde{Q} \in \mathbf{Q}$ such that $\tilde{P} = \tilde{Q}T^{-1}$ and $E_{\tilde{Q}}[g(R)I\{R \in \mathcal{R}'\}] = \theta_0\tilde{Q}\{R \in \mathcal{R}'\}$. By assumption, $\mathbf{P}_0$ is nonempty and therefore Lemma E.13 implies $\mathbf{P}_0(\tilde{P}) \equiv \{P \in \mathbf{P}_0 : P_Z = \tilde{P}_Z\}$ is nonempty as well, where P_Z denotes the marginal distribution of Z under P. Hence, there exists a $P \in \mathbf{P}_0$ such that $P_Z = \tilde{P}_Z$, $P = QT^{-1}$ for some $Q \in \mathbf{Q}$, $Q\{R \in \mathcal{R}'\} > 0$, and $\theta(Q) = \theta_0$. For a sequence $\lambda_n > 0$ satisfying $\lambda_n \downarrow 0$, define $P_n = \lambda_n P + (1-\lambda_n)\tilde{P}$ and $Q_n = \lambda_n Q + (1-\lambda_n)\tilde{Q}$. It then follows that $P_n = Q_n T^{-1}$, $Q_n\{R \in \mathcal{R}'\} > 0$, and $E_{Q_n}[g(R)I\{R \in \mathcal{R}'\}] = \theta_0 Q_n\{R \in \mathcal{R}'\}$, which implies that $\theta(Q_n) = \theta_0$. Moreover, since $Q = Q_R \times P_Z$ and $\tilde{Q} = \tilde{Q}_R \times P_Z$ for some distributions Q_R and $\tilde{Q}_R$, it follows that $Q_n = (\lambda_n Q_R + (1-\lambda_n)\tilde{Q}_R) \times P_Z$ and therefore that $R \perp\!\!\!\perp Z$ under Q_n. We conclude that $P_n \in \mathbf{P}_0$ for each n and, since $P_n \to \tilde{P}$ as $n \to \infty$, that $\tilde{P} \in \text{cl}(\mathbf{P}_0)$ implying $\tilde{\mathbf{P}}_0 \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$. ■

A.2 Proof of Theorem 4.2

We begin by introducing some notation that we use throughout the proofs. First let $\psi(V,P)$ be a vector of dimension $|\mathcal{M}|+2$. Indexing the first $|\mathcal{M}|$ coordinates of $\psi(V,P)$ by (y,d,z), we set

$$\psi_{ydz}(V,P) = \frac{1}{P_z} I\{Y=y, D=d, Z=z\} - \frac{P_{ydz}}{P_z^2} I\{Z=z\} \ , \tag{26}$$

and let the final two coordinates of $\psi(V,P)$ equal zero. The random variable $\psi(V,P)$ is the influence function of $\sqrt{n}\{\hat{\beta}_n - \beta(P)\}$ and will play an important role in our analysis.

The proof of Theorem 4.2 requires a final assumption that ensures that there is anti-concentration. To state this assumption, we let

$$\mathcal{V}(P) \equiv \{s \in \mathbf{R}^{|\mathcal{M}|+2} : A^{\dagger}s \leq 0, \|\Omega(P)(AA')^{\dagger}s\|_1 \leq 1\}$$

and let $\mathcal{E}(P)$ be the extreme points of $\mathcal{V}(P)$. For any $s \in \mathcal{E}(P)$, we also set $\sigma^2(s,P) \equiv s'\text{Var}_P\{\psi(s,P)\}s$ and define

$$\bar{\sigma}(P) \equiv \max_{s \in \mathcal{E}(P)} \sigma(s,P), \qquad \underline{\sigma}(P) \equiv \min_{s \in \mathcal{E}(P):\sigma(s,P)>0} \sigma(s,P) \ ,$$

where we set $\underline{\sigma}(P) = +\infty$ if $\sigma(s, P) = 0$ for all $s \in \mathcal{E}(P)$. Finally, for any $P \in \mathbf{P}$ we define

$$\mathrm{m}(P) = \mathrm{med}\left\{ \sup_{s \in \mathcal{V}(P)} \langle s, \mathbb{G}(P) \rangle \right\}$$

where for any random variable V, $\mathrm{med}\{V\}$ denotes its median. Given the introduced notation, we state a final assumption that ensures there is sufficient anti-concentration in our problem.

Assumption A.1. $\xi_n \equiv (\log^3(n)|\mathcal{M}|/n)^{1/4} \vee \lambda_n \sqrt{\log(|\mathcal{M}|)}$ satisfies $\sup_{P \in \mathbf{P}}(\mathrm{m}(P) + \bar{\sigma}(P))/\underline{\sigma}^2(P) = o(\xi_n^{-1})$.

We next provide the proof of Theorem 4.2. The proof relies on numerous auxiliary lemmas that are stated and proven in Online Appendix Section E.

Proof of Theorem 4.2. The proof follows the same arguments as those in the proof of Theorem 4.2 in Fang et al. (2023) but with their coupling results (Lemmas A.4 and A.5) replaced by our improved coupling rates from Lemma E.3. Specifically, we note that Lemma E.2 implies Assumptions 4.1, 4.2, 4.3, 4.4(v) and A.1 of Fang et al. (2023) hold. Assumptions 4.4(i)–(iv) are not verified directly, because their only role in the proof of Theorem 4.2 in Fang et al. (2023) is to ensure a coupling of the bootstrap statistic—a condition that we instead establish directly in Lemma E.3. The arguments in Fang et al. (2023) can therefore be applied with $r_n = \log^{3/2}(n)\sqrt{|\mathcal{M}|}/\sqrt{n}$ (the coupling rate of Lemma E.3(i)) and $b_n = r_n^{1/2}$ (the coupling rate of Lemma E.3(ii)). ∎

A.3 Proof of Theorem 5.1

Before proceeding with the proof, we first introduce some additional notation. Given the sets of restrictions $\mathcal{R} \subseteq \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$ and $\mathcal{R}' \subseteq \mathcal{Y}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|}$ we define, for any $r_t \in \mathcal{D}^{|\mathcal{Z}|}$, the sets

$$\mathcal{R}_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|D|} : (r_o, r_t) \in \mathcal{R}\}$$
$$\mathcal{R}'_o(r_t) \equiv \{r_o \in \mathcal{Y}^{|D|} : (r_o, r_t) \in \mathcal{R}'\} \ .$$

Recall $B_{c_o} \equiv B_{c_o(0),L} \times \cdots \times B_{c_o(|\mathcal{D}|-1),L} \subseteq \mathcal{Y}^{|\mathcal{D}|}$ for any $c_o \equiv (c_o(0), \ldots, c_o(|\mathcal{D}| - 1)) \in \mathcal{L}^{|\mathcal{D}|}$ and $\mathcal{M}_L \equiv \mathcal{L} \times \mathcal{D} \times \mathcal{Z}$. Let $\mathcal{N}_L \equiv \mathcal{L}^{|\mathcal{D}|} \times \mathcal{D}^{|\mathcal{Z}|} \times \{0, 1\}$. The matrix A_L then has $|\mathcal{M}_L| + 3$ rows and $|\mathcal{N}_L| + 2$ columns.

To construct A_L, we index the first $|\mathcal{M}_L|$ rows of A_L by $(\ell, d, z) \in \mathcal{M}_L$ and the first $|\mathcal{N}_L|$ columns by $r = (c_o, r_t, \chi)$ where $c_o = (c_o(d) : d \in \mathcal{D})$, $r_t = (d(z) : z \in \mathcal{Z})$, and $\chi \in \{0, 1\}$. Define

$$\mathcal{Y}(c_o, r_t, \chi) = \begin{cases} aB_{c_o} \cap \mathcal{R}'_o(r_t) & \text{if } \chi = 1 \\ B_{c_o} \backslash \mathcal{R}'_o(r_t) & \text{if } \chi = 0 \ . \end{cases}$$

If $\mathcal{Y}(r) = \emptyset$, then the r column of A_L is zero. Otherwise, the $(\ell, d, z) \times r$ element of A_L is given by $I\{c_o(d) = \ell, d(z) = d\}$ and the $(|\mathcal{M}_L| + 1) \times r$ element of A_L is given by $I\{(\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}) \cap \mathcal{R} \neq \emptyset\}$. We also let the $(|\mathcal{M}_L| + 2) \times r$ and $(|\mathcal{M}_L| + 3) \times r$ elements of A_L be respectively given by

$$\left(\sup_{r_o \in \mathcal{Y}(c_o, r_t, \chi)} g(r_o, r_t) - \theta_0 \right) \times \chi$$
$$\left(\inf_{r_o \in \mathcal{Y}(c_o, r_t, \chi)} g(r_o, r_t) - \theta_0 \right) \times \chi \ .$$

Finally, the $(|\mathcal{M}_L|+2) \times (|\mathcal{N}_L|+1)$ element of A_L is given by -1 and the $(|\mathcal{M}_L|+3) \times (|\mathcal{N}_L|+2)$ element of A_L is given by 1. All other elements of A_L are given by 0.

Given the introduced definitions, we next prove Theorem 5.1.

PROOF OF THEOREM 5.1. First suppose $P \in \mathbf{P}_0$. Then, there is a $Q \in \mathbf{Q}_0(P, \mathbf{Q})$ such that $\theta_0 = \theta(Q)$ and $Q\{R \in \mathcal{R}'\} > 0$. Letting $q(c_o, r_t, \chi) \equiv Q\{R_o \in \mathcal{Y}(c_o, r_t, \chi), R_t = r_t\}$ for any $c_o \in \mathcal{L}^{|\mathcal{D}|}$ and $r_t \in \mathcal{D}^{|\mathcal{Z}|}$, we then obtain

$$P_{\ell d|z} = Q\{Y(d) \in B_\ell, D(z) = d\} = \sum_{\substack{c_o \in \mathcal{L}^{|\mathcal{D}|}, r_t \in \mathcal{R}_t \\ \chi \in \{0,1\}}} q(c_o, r_t, \chi) I\{c_o(d) = \ell, r_t(z) = d\} \ . \qquad (27)$$

Collect the probabilities $q(c_o, r_t, \chi)$ into a vector $q = (q(c_o, r_t, \chi) : c_o \in \mathcal{L}^{|\mathcal{D}|}, r_t \in \mathcal{D}^{|\mathcal{Z}|}, \chi \in \{0, 1\})$, and define ξ_1, ξ_2 by

$$\xi_1 \equiv \sum_{c_o \in \mathcal{L}^{|\mathcal{D}|}, r_t \in \mathcal{R}_t} \left(\sup_{r_o \in B_{c_o} \cap \mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0 \right) q(c_o, r_t, 1)$$
$$\xi_2 \equiv - \sum_{c_o \in \mathcal{L}^{|\mathcal{D}|}, r_t \in \mathcal{R}_t} \left(\inf_{r_o \in B_{c_o} \cap \mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0 \right) q(c_o, r_t, 1) \ .$$

Next, note that since $\theta(Q) = \theta_0$ by hypothesis and $Q\{R \in \mathcal{R}'\} > 0$, it follows that Q satisfies

$E_Q[(g(R)-\theta_0)I\{R\in\mathcal{R}'\}]=0$. In particular, by definition of ξ_1 and ξ_2 we then obtain that

$$-\xi_1 \leq E_Q[(g(R)-\theta_0)I\{R\in\mathcal{R}'\}]=0\leq \xi_2 \ .$$

Setting $x=(q',\xi_1,\xi_2)'$ we then note that $x\geq 0$. Moreover, it follows from (27) and the definition of A_L that the first $|\mathcal{M}_L|$ rows of $A_L x$ equal the first $|\mathcal{M}_L|$ rows of $\beta(P)$. Also note that because $Q\in\mathbf{Q}_0(P,\mathbf{Q})$ implies Q satisfies Assumption 2.2 and hence $Q\{R\in\mathcal{R}\}=1$ it follows that

$$\sum_{\substack{c_o\in\mathcal{L}^{|\mathcal{D}|},r_t\in\mathcal{R}_t\\ \chi\in\{0,1\}}} q(c_o,r_t,\chi)\times I\{(\mathcal{Y}(c_o,r_t,\chi)\times\{r_t\})\cap\mathcal{R}\neq\emptyset\}=1\ ,$$

and therefore the $|\mathcal{M}_L|+1$ row of $A_L x$ equals the $|\mathcal{M}_L|+1$ row of $\beta(P)$ by definition of A_L and $\beta(P)$. Finally, we note the $|\mathcal{M}_L|+2$ and $|\mathcal{M}_L|+3$ rows of $A_L x$ equal the $|\mathcal{M}_L|+2$ and $|\mathcal{M}_L|+3$ rows of $\beta_L(P)$ by definition of ξ_1, ξ_2, A_L and $\beta_L(P)$. Therefore, we have shown that $\mathbf{P}_0\subseteq\mathbf{P}_{0,L}\equiv\{P\in\mathbf{P}: Ax=\beta(P)\text{ for some } x\geq 0\}$.

It remains to show that $\limsup_{L\to\infty}\mathbf{P}_{0,L}\subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$. We proceed by contradiction, and assume that there is a $P\in\limsup_{L\to\infty}\mathbf{P}_{0,L}$ such that $P\notin\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$. Since $P\notin\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$, Lemma E.12 implies that

$$\int f dP=1 \qquad\qquad \int f dP'\leq 0 \text{ for all } P'\in\mathbf{P}_0 \text{ satisfying } P'_Z=P_Z \tag{28}$$

for some continuous and bounded function $f:\mathcal{M}\to\mathbf{R}$ and P_Z denoting the marginal distribution of Z under P. Since $f:\mathcal{M}\to\mathbf{R}$ is continuous and $\mathcal{Y}$ is compact by Assumption 5.1(i), it follows that f is uniformly continuous on $\mathcal{M}$. For any $\eta>0$ there therefore exists a $\delta>0$ such that

$$\max_{d\in\mathcal{D},z\in\mathcal{Z}}\sup_{|y-y'|\leq 2\delta}|f(y,d,z)-f(y',d,z)|\leq\eta\ . \tag{29}$$

Moreover, by Assumption 5.1(iii) there is a $L_0<\infty$ such that $\max_{1\leq\ell\leq L_0}\text{diam}\{B_{\ell,L}\}<\delta$ for all $L\geq L_0$. Since $P\in\limsup_{L\to\infty}\mathbf{P}_{0,L}$ implies $P\in\mathbf{P}_{0,L}$ for some $L\geq L_0$, we conclude there is some $L<\infty$ such that

$$P\in\mathbf{P}_{0,L} \quad\text{ and }\quad \max_{1\leq\ell\leq L}\text{diam}\{B_{\ell,L}\}<\delta\ . \tag{30}$$

Next define λ to be a vector of dimension $|\mathcal{M}_L|+3$ whose last three entries are zero, and each of its first $|\mathcal{M}_L|$ entries, which we index by $(\ell, d, z) \in \mathcal{L} \times \mathcal{D} \times \mathcal{Z}$, are given by

$$\lambda(\ell, d, z) \equiv \sup_{y \in B_{\ell,L}} f(y, d, z) P\{Z = z\} .$$

Also observe that $P \in \mathbf{P}_{0,L}$ implies $\beta_L(P) = A_L x$ for some $x \geq 0$. Indexing the first $|\mathcal{N}_L|$ entries of x by $r = (c_o, r_t, \chi)$ with $(c_o, r_t, \chi) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t \times \{0, 1\}$, let Q'_R be any distribution of R satisfying $Q'_R\{R_o \in \mathcal{Y}(c_o, r_t, \chi), R_t = r_t\} = x(c_o, r_t, \chi)$. Defining the set of indices $\mathcal{I}(\mathcal{R}) \equiv \{(c_o, r_t, \chi) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t : \mathcal{Y}(c_o, r_t, \chi) \cap \mathcal{R} \neq \emptyset\}$ and then note that the last three entries of $A_L x$ being equal to $(1, 0, 0)$ because $A_L x = \beta_L(P)$ imply that the distribution Q'_R satisfies

$$\sum_{(c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R})} Q'_R\{(R_o, R_t) \in (\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\})\} = 1$$

$$\sum_{(c_o, r_t) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t} \left(\sup_{r_o \in B_{c_o} \cap \mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0 \right) \times Q'_R\{(R_o, R_t) \in (B_{c_o} \times \{r_t\}) \cap \mathcal{R}'\} \geq 0$$

$$\sum_{(c_o, r_t) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t} \left(\inf_{r_o \in B_{c_o} \cap \mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0 \right) \times Q'_R\{(R_o, R_t) \in (B_{c_o} \times \{r_t\}) \cap \mathcal{R}'\} \leq 0 . \quad (31)$$

Setting $Q' \equiv Q'_R \times P_Z$ and letting $P' = Q' T^{-1}$, then note that by construction we have $\beta_L(P) = \beta_L(P')$. Therefore, using that f satisfies (29), $\max_{1 \leq \ell \leq L} \operatorname{diam}\{B_{\ell,L}\} < \delta$ by (30), and the definition of λ gives us

$$\int f dP \leq \sum_{(\ell, d, z) \in \mathcal{M}_L} P\{Y \in B_{\ell,L}, D = d, Z = z\} \times \sup_{y \in B_{\ell,L}} f(y, d, z)$$
$$= \lambda' \beta_L(P) = \lambda' \beta_L(P') \leq \int f dP' + \eta . \quad (32)$$

On the other hand, Q'_R satisfying (31) implies by Lemma E.11 that there is a distribution $\tilde{Q}_R$ for R such that $\tilde{Q} = \tilde{Q}_R \times P_Z$ satisfies $E_{\tilde{Q}}[(g(R) - \theta_0) I\{R \in \mathcal{R}'\}]$, $\tilde{Q}\{R \in \mathcal{R}\} = 1$, and such that $\tilde{P} = \tilde{Q} T^{-1}$ satisfies

$$\int f(dP' - d\tilde{P}) \leq 2\eta . \quad (33)$$

Finally, let $\mathbf{P}_0(P) \equiv \{\hat{P} \in \mathbf{P}_0 : \hat{P}_Z = P_Z\}$ and note that Lemma E.13 implies $\mathbf{P}_0(P)$ is not empty. Pick any $P_0 \in \mathbf{P}_0(P)$ and note that $P_0 = Q_0 T^{-1}$ for some $Q_0 \in \mathbf{Q}$ satisfying $Q_0 = Q_{0,R} \times P_Z$ for some distribution $Q_{0,R}$ for R. For any $\tau \in [0, 1]$, it then follows that $Q_\tau \equiv (1 - \tau)\tilde{Q} + \tau Q_0 = \{(1 - \tau)\tilde{Q}_R + \tau Q_{0,R}\} \times P_Z$ satisfies $E_{Q_\tau}[(g(R) - \theta_0) I\{R \in \mathcal{R}'\}] = 0$, $Q_\tau\{R \in \mathcal{R}\} = 1$, and $Q_\tau\{R \in \mathcal{R}'\} > 0$ because $Q_0\{R \in \mathcal{R}'\} > 0$. In particular, it follows that $P_\tau \equiv Q_\tau T^{-1} \in \mathbf{P}_0(P)$. Moreover, since $f : \mathcal{M} \to \mathbf{R}$ is bounded, we also have that

$|\int f(d\tilde{P} - dP_\tau)| \leq 2\tau\|f\|_\infty$. Therefore, setting $\tau \leq \eta/(2\|f\|_\infty)$ allows us to conclude that

$$\int f d\tilde{P} \leq \int f dP_\tau + \eta \leq \eta \ , \tag{34}$$

where the final inequality holds by (28) and $P_\tau \in \mathbf{P}_0(P)$. Combining results (32), (33), and (34) finally yield $\int f dP \leq 4\eta$, which contradicts (28) for η sufficiently small. Hence, we conclude that there is no $P \in \limsup_{L\to\infty} \mathbf{P}_{0,L}$ satisfying $P \notin \mathrm{cl}_{\mathbf{P}}(\mathbf{P}_0)$, which completes the proof. ■

B Details for Remark 4.1

In this section, we describe a method for deriving analytical expressions for $\Theta_0(P, \mathbf{Q})$ as an interval whose lower and upper endpoints are some functions of P, $L(P)$, and $U(P)$ respectively. If Q satisfies Assumption 2.1 and $P = QT^{-1}$, then $P\{Z = z\} = Q\{Z = z\}$ and (24) holds. Therefore, in what follows, we disregard the distribution of Z under Q and identify Q with the distribution of R, and in turn with the column vector q. Correspondingly, we identify P with $p = \{p_{yd|z} : (y, d, z) \in \mathcal{M}\}$. Let A_0 and $\beta_0(P)$ denote the first $|\mathcal{M}| + 1$ rows of A and $\beta(P)$. Because we assume every $Q \in \mathbf{Q}$ satisfies Assumption 2.1,

$$\mathbf{Q}_0(P, \mathbf{Q}) = \{q : A_0 q = \beta_0(P),\ q \geq 0\} \ . \tag{35}$$

B.1 When $Q\{R \in \mathcal{R}'\}$ is known or identified

Suppose $Q\{R \in \mathcal{R}'\} > 0$ for all $Q \in \mathbf{Q}$ and is known or identified, so that it is a function of P for all P such that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. Denote such a function by $a(P)$. Recall q introduced in the proof of Theorem 4.1, which is indexed by r such that $q(r) = Q\{R = r\}$. For all $Q \in \mathbf{Q}_0(P, \mathbf{Q})$, $\theta(Q) = \frac{1}{a(P)} \sum_{r \in \mathcal{R}'} g(r) q(r)$. Let c be a column vector indexed by r, with the r element given by $g(r)/a(P)$. Here we suppress the dependence of c on P for convenience. With the notation above,

$$\Theta_0(P, \mathbf{Q}) = c' \mathbf{Q}_0(P, \mathbf{Q}) \ .$$

Before proceeding, we show $\Theta_0(P, \mathbf{Q})$ is a closed interval. Because $\mathbf{Q}_0(P, \mathbf{Q})$ is a polyhedron (in standard form), Corollary 2.5 in Bertsimas and Tsitsiklis (1997) implies $\Theta_0(P, \mathbf{Q}) = c' \mathbf{Q}_0(P, \mathbf{Q})$ is a polyhedron in $\mathbf{R}$. Furthermore, $\mathbf{Q}_0(P, \mathbf{Q})$ is obviously bounded, so $\Theta_0(P, \mathbf{Q})$ is also bounded. A bounded polyhedron in $\mathbf{R}$ is simply the intersection of bounded closed intervals, so it is itself a bounded closed interval.

Next, note $L(P)$ is the solution to

$$\begin{aligned} \min_q \quad & c'q \\ \text{subject to} \quad & A_0 q = \beta_0(P) \\ & q \geq 0 \ . \end{aligned} \tag{36}$$

Because the feasible set $\mathbf{Q}_0(P, \mathbf{Q})$ in (36) is in addition bounded, the optimal cost has to be finite, so it follows from Theorem 2.8 in Bertsimas and Tsitsiklis (1997) that (36) has an optimal solution. Therefore, by the strong duality theorem (Theorem 4.4, Bertsimas and Tsitsiklis, 1997), the optimal value in (36) is the same as the optimal value of its dual:

$$\begin{aligned} \max_r \quad & \beta_0(P)'r \\ \text{subject to} \quad & A_0' r \leq c \ . \end{aligned} \tag{37}$$

Note that in (37) the constraints do not involve P. It follows from the resolution theorem (Exercise 4.47, Bertsimas and Tsitsiklis, 1997) that the feasible set in (37) can be written as

$$\left\{ \sum_{1 \leq j \leq J} \theta_j r_j^{\text{ex}} + \sum_{1 \leq \ell \leq L} \lambda_\ell r_\ell^{\text{ray}} : \theta_j \geq 0, \sum_{1 \leq j \leq J} \theta_j = 1, \lambda_\ell \geq 0 \right\} \ ,$$

where each r_j^{ex} is a vertex and each r_ℓ^{ray} spans an extreme ray. Because (37) cannot be unbounded, $b' r_\ell^{\text{ray}} \leq 0$ for each ℓ, so the optimal solution of (37) must have $\lambda_\ell = 0$ for all ℓ. Therefore, it follows from Theorem 2.8 of Bertsimas and Tsitsiklis (1997) that the optimal value for (37) is

$$\max_{1 \leq j \leq J} b' r_j^{\text{ex}} \ .$$

The conclusion follows because the class of vertices r_j^{ex} are determined by A_0 and c and do not depend on P. $U(P)$ can be obtained in a similar fashion.

B.2 When $Q\{R \in \mathcal{R}'\}$ is partially identified

Suppose now $Q\{R \in \mathcal{R}'\}$ is not point identified. In this case, $\Theta_0(P, \mathbf{Q})$ is not necessarily a closed interval, since the set $\{Q : Q \in \mathbf{Q}_0(P, \mathbf{Q}),\ Q\{R \in \mathcal{R}'\} > 0\}$ is not necessarily closed, but the latter remains a connected set. Recall (25) that

$$\theta(Q) = \frac{E_Q[g(R) I\{R \in \mathcal{R}'\}]}{Q\{R \in \mathcal{R}'\}} \ ,$$

so $\theta(Q)$ is a continuous function of Q whenever $Q\{R \in \mathcal{R}'\} > 0$. As a result, the identified set $\Theta_0(P, \mathbf{Q}) \equiv \{\theta(Q) : Q \in \mathbf{Q}_0(P, \mathbf{Q}),\ Q\{R \in \mathcal{R}'\} > 0\}$ as the image of a continuous function on a connected set is also connected. Since $\Theta_0(P, \mathbf{Q}) \subseteq \mathbf{R}$, it is an interval (though not necessarily closed). Its lower endpoint, $L(P)$, can be obtained by solving the optimization problem

$$L(P) = \inf_{\substack{Q \in \mathbf{Q}_0(P,\mathbf{Q}) \\ Q\{R \in \mathcal{R}'\} > 0}} \frac{E_Q[g(R) I\{R \in \mathcal{R}'\}]}{Q\{R \in \mathcal{R}'\}} ,$$

and similarly for the upper endpoint.

The optimization problem above is known as a linear fractional program and can be transformed into a linear program using the Charnes–Cooper transformation (Charnes and Cooper, 1962). One may then be tempted to replicate the strategy in Section B.1 relying on the linear program from the Charnes–Cooper transformation. Such a strategy generally fails because in the dual program of the transformed linear program, in contrast with (37), the constraints are not known ex-ante but are functions of P. As a result, we propose an alternative strategy as follows.

The problem of finding the analytical expressions for $L(P)$ (or $U(P)$) can be written as the following two-step optimization problem which is easier to analyze:

$$\begin{aligned} \inf_{\substack{Q \in \mathbf{Q}_0(P,\mathbf{Q}) \\ Q\{R \in \mathcal{R}'\} > 0}} \frac{E_Q[g(R) I\{R \in \mathcal{R}'\}]}{Q\{R \in \mathcal{R}'\}} &\iff \inf_{\pi \in \Pi(\mathcal{R}', P, \mathbf{Q}) \setminus \{0\}} \left\{ \inf_{Q \in \mathbf{Q}_0(P,\mathbf{Q}) \cap \Delta(\pi)} \frac{E_Q[g(R) I\{R \in \mathcal{R}'\}]}{\pi} \right\} \\ &\iff \inf_{\pi \in \Pi(\mathcal{R}', P, \mathbf{Q}) \setminus \{0\}} \frac{1}{\pi} \left\{ \inf_{Q \in \mathbf{Q}_0(P,\mathbf{Q}) \cap \Delta(\pi)} E_Q[g(R) I\{R \in \mathcal{R}'\}] \right\} , \end{aligned} \tag{38}$$

where for any $\pi \in [0, 1]$, we define

$$\Delta(\pi) \equiv \{Q \in \mathbf{Q} : \pi = Q\{R \in \mathcal{R}'\}\} , \tag{39}$$

and

$$\Pi(\mathcal{R}', P, \mathbf{Q}) \equiv \{Q\{R \in \mathcal{R}'\} : Q \in \mathbf{Q}_0(P, \mathbf{Q})\} , \tag{40}$$

as the identified set for $Q\{R \in \mathcal{R}'\}$. Solving for $\Pi(\mathcal{R}', P, \mathbf{Q})$, however, is a special case of Section B.1 by specifying $\mathcal{R}' = \mathcal{R}$ and $g(R) = I\{R \in \mathcal{R}'\}$ in $\theta(Q) = E_Q[g(R) | R \in \mathcal{R}']$, so that $Q\{R \in \mathcal{R}'\} = 1$ by Assumption 2.2. Therefore, $\Pi(\mathcal{R}', P, \mathbf{Q})$ is a closed interval whose lower (resp. upper) endpoint is the maximum (resp. minimum) of a finite number of linear functions of $(p_{yd|z}(P))$.

For the inner minimization problem, $\mathbf{Q}_0(P, \mathbf{Q}) \cap \Delta(\pi)$ is characterized by

$$\mathbf{Q}_0(P, \mathbf{Q}) \cap \Delta(\pi) = \{q : A(\mathcal{R}')q = \beta(P, \pi),\ q \geq 0\}\ ,$$

where $A(\mathcal{R}')$ is a $(|\mathcal{M}| + 2) \times |\mathcal{N}|$ matrix, whose first $|\mathcal{M}| + 1$ rows are the same as A_0, and the last row is given by $I\{r \in \mathcal{R}'\}$, and $\beta(P, \pi) \equiv (p_{yd|z}(P) : (y, d, z) \in \mathcal{M}, 1, \pi)'$. Using the same analysis in Section B.1,

$$\inf_{Q \in \mathbf{Q}_0(P,\mathbf{Q}) \cap \Delta(\pi)} E_Q[g(R)I\{R \in \mathcal{R}'\}]$$

is given by the maximum of a finite number of linear functions of $(p_{yd|z}(P)) \cup (\pi)$, denoted as $L(P, \pi)$. As a result, $L(P)$ has the following form:

$$L(P) = \inf_{\pi \in \Pi(\mathcal{R}', P, \mathbf{Q}) \setminus \{0\}} \frac{1}{\pi} L(P, \pi)\ , \tag{41}$$

and $U(P)$ can be obtained in a similar fashion.

C Details for Remark 4.2

Consider the setting of Example 3.1, an RCT with one-sided noncompliance, and additionally suppose $\mathcal{D} = \mathcal{Z} = \{0, 1, 2\}$ and $\mathcal{Y} = \{0, 1\}$. Let $\pi_{0ij} = Q\{R^t = (0, i, j)\}$ for $i \in \{0, 1\}$ and $j \in \{0, 2\}$. We now show that the identified set on π_{012} depends on the distribution of the full vector (Y, D, Z) instead of only the distribution of (D, Z) by showing that the bounds on π_{012} that only use the distribution of (D, Z) differ from the identified set on π_{012}.

First, consider the bounds on π_{012} that only use the distribution of (D, Z). Cheng and Small (2006) construct bounds on π_{012} in this setting by minimizing and maximizing π_{012} under the following constraints:

$$\begin{aligned}
p_{1|1} &= \pi_{012} + \pi_{010} \\
p_{0|1} &= \pi_{002} + \pi_{000} \\
p_{2|2} &= \pi_{002} + \pi_{012} \\
p_{0|2} &= \pi_{000} + \pi_{010} \\
1 &= \pi_{012} + \pi_{010} + \pi_{002} + \pi_{000} \\
0 \leq \pi_{012},\ \pi_{010},&\ \pi_{002},\ \pi_{000} \leq 1\ ,
\end{aligned} \tag{42}$$

where $p_{d|z} = P\{D = d|Z = Z\}$ for $d \in \mathcal{D}$ and $z \in \mathcal{Z}$. Note that $\{Q \in \mathbf{Q} : Q$ satisfies (42)$\}$

is the set of all $Q \in \mathbf{Q}$ that rationalize $\{p_{d|z} : d \in \mathcal{D}, z \in \mathcal{Z}\}$. Forming bounds on π_{012} by this linear program is the same as the procedure outlined in Appendix B.1, except that it only uses as constraints that Q rationalizes $\{p_{d|z} : d \in \mathcal{D}, z \in \mathcal{Z}\}$. By parallel arguments to those of Appendix B.1, these bounds correspond to the analog to the identified set if defined only using the marginal distribution of (D, Z) as opposed to the full vector (Y, D, Z). Let $\Theta_m(P, \mathbf{Q})$ denote the resulting bounds on π_{012}, which Cheng and Small (2006) show is given by

$$\Theta_m(P, \mathbf{Q}) = \left[\max\{0,\ p_{1|1} - p_{0|2}\},\ \min\{p_{1|1},\ p_{2|2}\} \right] . \tag{43}$$

In contrast, applying our procedure from Appendix B that uses as constraints that Q rationalizes $\{p_{yd|z} : y \in \mathcal{Y}, d \in \mathcal{D}, z \in \mathcal{Z}\}$ results in the identified set on π_{012} being given by

$$\begin{aligned}\Theta_0(P, \mathbf{Q}) &= \left[\max \left\{ \begin{array}{c} 0 \\ p_{01|1} + p_{11|1} - (p_{00|2} + p_{10|2}) \\ p_{10|0} - p_{10|1} - p_{10|2} \\ 1 - p_{00|1} - p_{00|2} - p_{10|0} \end{array} \right\} ,\ \min \left\{ \begin{array}{c} p_{01|1} + p_{11|1} \\ p_{02|2} + p_{12|2} \\ 1 - p_{00|1} - p_{10|2} \\ 1 - p_{10|1} - p_{00|2} \end{array} \right\} \right] \\ &= \left[\max \left\{ \begin{array}{c} 0 \\ p_{1|1} - p_{0|2} \\ p_{10|0} - p_{10|1} - p_{10|2} \\ 1 - p_{00|1} - p_{00|2} - p_{10|0} \end{array} \right\} ,\ \min \left\{ \begin{array}{c} p_{1|1} \\ p_{2|2} \\ 1 - p_{00|1} - p_{10|2} \\ 1 - p_{10|1} - p_{00|2} \end{array} \right\} \right] ,\end{aligned} \tag{44}$$

Under Assumption 2.1 and Assumption 2.2 with $\mathcal{R}$ defined as in Example 3.1, we can rewrite the right-hand side of (43) in terms of Q as

$$\left[\max\{0,\ Q\{R^t = (0,1,2)\} - Q\{R^t = (0,0,0)\}\},\ \min\{Q\{D_1 = 1\},\ Q\{D_2 = 2\}\} \right] , \tag{45}$$

and the right-hand side of (44) as

$$\begin{aligned}&\left[\max \left\{ \begin{array}{c} 0 \\ Q\{R^t = (0,1,2)\} - Q\{R^t = (0,0,0)\} \\ Q\{Y(0) = 1, R^t = (0,1,2)\} - Q\{Y(0) = 1, R^t = (0,0,0)\} \\ Q\{Y(0) = 0, R^t = (0,1,2)\} - Q\{Y(0) = 0, R^t = (0,0,0)\} \end{array} \right\} , \right. \\ &\qquad \left. \min \left\{ \begin{array}{c} Q\{D_1 = 1\} \\ Q\{D_2 = 2\} \\ Q\{Y(0) = 0, D_1 = 1\} + Q\{Y(0) = 1, D_2 = 2\} \\ Q\{Y(0) = 1, D_1 = 1\} + Q\{Y(0) = 0, D_2 = 2\} \end{array} \right\} \right] .\end{aligned} \tag{46}$$

These expressions make it transparent why both (43) and (44) are valid bounds on $\pi_{012} = Q\{R^t = (0,1,2)\}$ under the maintained assumptions and how (44) incorporates information from the distribution of outcomes.

We now consider when (45) will strictly contain (46). The lower bound of (46) will be strictly greater than the lower bound of (45) if

$$\frac{Q\{Y(0) = j \mid R^t = (0,1,2)\}}{Q\{Y(0) = j \mid R^t = (0,0,0)\}} < \frac{Q\{R^t = (0,0,0)\}}{Q\{R^t = (0,1,2)\}} < \frac{Q\{Y(0) = 1-j \mid R^t = (0,1,2)\}}{Q\{Y(0) = 1-j \mid R^t = (0,0,0)\}} \tag{47}$$

for $j = 0$ or 1, and otherwise the lower bounds coincide. Likewise, the upper bound of (46) will be strictly smaller than the upper bound of (45) if

$$\frac{Q\{Y(0) = j \mid D_1 = 1\}}{Q\{Y(0) = j \mid D_2 = 2\}} < \frac{Q\{D_2 = 2\}}{Q\{D_1 = 1\}} < \frac{Q\{Y(0) = 1-j \mid D_1 = 1\}}{Q\{Y(0) = 1-j \mid D_2 = 2\}} \tag{48}$$

for $j = 0$ or 1, and otherwise the upper bounds coincide. We conclude that (45) and (46) will coincide if the $Y(0)$ potential outcome is independent of treatment responses, while the bounds of (46) will be strictly smaller than that of (45) if the dependence between $Y(0)$ and the treatment response types is sufficiently strong. Thus, the bounds on π_{012} of (43) will strictly contain the bounds of (44) for any P such that $P = QT^{-1}$ for a Q with sufficient dependence between $Y(0)$ and treatment response types. This role of dependence between potential outcomes and treatment responses is reminiscent of the role of such dependence in the ability to detect violations of latent monotonicity restrictions shown in Machado et al. (2019).

We now consider a numerical example with the Q distribution specified in Table 5 and the implied P distribution specified in Table 4, where we write $q(y_0y_1y_2, d_0d_1d_2) = Q\{Y(d) = y_d, D(z) = d_z,\ (d,z) \in \mathcal{D} \times \mathcal{Z}\}$ and omit any $q(\cdot) = 0$. One can check that $Q \in \mathbf{Q}$ and $P = QT^{-1}$ so that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$. Moreover, $Q\{R \in \mathcal{R}\} = 1$ for $\mathcal{R}$ defined in Example 3.1. In this numerical example, $\Theta_m(P, \mathbf{Q}) = [0.195, 0.481]$, while $\Theta_0(P, \mathbf{Q}) = [0.235, 0.419]$. Thus, in this numerical example, P is such that $\mathbf{Q}_0(P, \mathbf{Q}) \neq \emptyset$ with $\Theta_m(P, \mathbf{Q}) \supsetneq \Theta_0(P, \mathbf{Q})$. Note that, in this example,

$$\begin{aligned}
\frac{Q\{Y(0) = 1 \mid R^t = (0,1,2)\}}{Q\{Y(0) = 1 \mid R^t = (0,0,0)\}} &= 0.165 \\
\frac{Q\{R^t = (0,0,0)\}}{Q\{R^t = (0,1,2)\}} &= 0.464 \\
\frac{Q\{Y(0) = 0 \mid R^t = (0,1,2)\}}{Q\{Y(0) = 0 \mid R^t = (0,0,0)\}} &= 1.49
\end{aligned}$$

so that the strong dependence between $Y(0)$ and treatment response types causes (47) to hold for $j = 1$ and thus the lower bound of $\mathbf{Q}_m(P, \mathbf{Q})$ to be strictly larger than the lower bound of $\mathbf{Q}_0(P, \mathbf{Q})$. Likewise,

$$\frac{Q\{Y(0) = 0 \mid D_1 = 1\}}{Q\{Y(0) = 0 \mid D_2 = 2\}} = 0.544$$
$$\frac{Q\{D_2 = 2\}}{Q\{D_1 = 1\}} = 1.48$$
$$\frac{Q\{Y(0) = 1 \mid D_1 = 1\}}{Q\{Y(0) = 1 \mid D_2 = 2\}} = 1.93$$

so that the strong dependence between $Y(0)$ and treatment response types causes (48) to hold for $j = 0$ and thus the upper bound of $\mathbf{Q}_0(P, \mathbf{Q})$ to be strictly smaller than the upper bound of $\mathbf{Q}_m(P, \mathbf{Q})$.

	$p_{00\|0}$ 0.764	$p_{10\|0}$ 0.236	
$p_{00\|1}$ 0.412	$p_{10\|1}$ 0.107	$p_{01\|1}$ 0.301	$p_{11\|1}$ 0.180
$p_{00\|2}$ 0.117	$p_{10\|2}$ 0.169	$p_{02\|2}$ 0.475	$p_{12\|2}$ 0.239

Table 4: Distribution P for Appendix C.

$q(000, 000)$ 0.002	$q(000, 010)$ 0.002	$q(000, 002)$ 0.017	$q(000, 012)$ 0.025
$q(001, 000)$ 0.002	$q(001, 010)$ 0.002	$q(001, 002)$ 0.002	$q(001, 012)$ 0.195
$q(010, 000)$ 0.101	$q(010, 010)$ 0.002	$q(010, 002)$ 0.272	$q(010, 012)$ 0.120
$q(011, 000)$ 0.002	$q(011, 010)$ 0.004	$q(011, 002)$ 0.014	$q(011, 012)$ 0.002
$q(100, 000)$ 0.002	$q(100, 010)$ 0.002	$q(100, 002)$ 0.022	$q(100, 012)$ 0.011
$q(101, 000)$ 0.034	$q(101, 010)$ 0.062	$q(101, 002)$ 0.015	$q(101, 012)$ 0.002
$q(110, 000)$ 0.002	$q(110, 010)$ 0.002	$q(110, 002)$ 0.006	$q(110, 012)$ 0.002
$q(111, 000)$ 0.024	$q(111, 010)$ 0.041	$q(111, 002)$ 0.002	$q(111, 012)$ 0.007

Table 5: Distribution Q for Appendix C.

D Details for Testable Implications

In this section, we discuss a method to obtain sharp testable restrictions of $\mathbf{Q}$ in terms of analytical inequalities. In what follows, we will make heavy use of the fact that a nonempty bounded polyhedron can be represented in two ways. Recall from Definition 2.1 of Bertsimas and Tsitsiklis (1997) that a polyhedron in $\mathbf{R}^k$ is a set $\{x \in \mathbf{R}^k : Ax \leq b\}$, known as the H-representation of a polyhedron. If the polyhedron is nonempty and bounded, then Theorem 2.9 of Bertsimas and Tsitsiklis (1997) implies that it can equivalently be represented as the convex hull of its (finite number of) vertices, known as the V-representation of a polyhedron.

As in Appendix B, we identify Q with the column vector $q = (q(r) : r \in \mathcal{R})$, and P with $p = \{p_{yd|z} : (y, d, z) \in \mathcal{M}\}$. Further let A_1 denote the first $|\mathcal{M}|$ rows of A_0 and let a_0 denote the last row of A_0. Correspondingly, we note

$$\mathbf{Q} = \{q : a_0' q = 1, q \geq 0\} ,$$

which is a bounded polyhedron in H-representation. Next, let $\mathbf{P}(\mathbf{Q}) = \{P : P = QT^{-1}, Q \in \mathbf{Q}\}$, and note $\mathbf{P}(\mathbf{Q}) = A_1\mathbf{Q}$. $\mathbf{P}(\mathbf{Q})$ is obviously a bounded polyhedron and is nonempty as long as $\mathbf{Q}$ is nonempty. In that case, Theorem 2.9 of Bertsimas and Tsitsiklis (1997) implies it is the convex hull of its (finite number of) vertices.

The previous discussion leads to the following algorithm for obtaining $\mathbf{P}(\mathbf{Q})$ in terms of inequalities, i.e., its H-representation. For a given polyhedron, we can compute one representation from the other using the `mpt3` package in `MATLAB`. To obtain the H-representation of $\mathbf{P}(\mathbf{Q})$, we use the following algorithm:

Algorithm D.1.

Step 1: Collect the set of all vertices of $\mathbf{Q}$, denoted by $V = \{V_i : 1 \leq i \leq n\}$.

Step 2: Compute $A_1 V = \{A_1 V_i : 1 \leq i \leq n\}$.

Step 3: From Lemma D.1, define $A_1\mathbf{Q} = \text{co}(A_1 V)$.

Step 4: Obtain the H representation of $A_1\mathbf{Q}$.

In the algorithm, we have used the following lemma that allows us to define $\mathbf{P}(\mathbf{Q})$ through the vertices of $\mathbf{Q}$:

Lemma D.1. *Suppose $\mathbf{Q}$ is a non-empty bounded polyhedron and V is the set of vertices of $\mathbf{Q}'$. Then, $A_1\mathbf{Q} = \text{co}(A_1 V)$.*

Proof. We first show $\mathrm{co}(A_1V) \subseteq A_1\mathbf{Q}'$. Indeed, each $P \in \mathrm{co}(A_1V)$ could be written as $\sum_{1\leq i\leq n} \lambda_i A_1 V_i = A_1 \sum_{1\leq i\leq n} \lambda_i V_i$ where $\lambda_i \geq 0$ for all i and $\sum_i \lambda_i = 1$, but $\sum_{1\leq i\leq n} \lambda_i V_i \in \mathbf{Q}$ since $\mathbf{Q}$ is a polyhedron and hence convex. To show $A_1\mathbf{Q} \subseteq \mathrm{co}(A_1V)$, fix $Q \in \mathbf{Q}$. Because $\mathbf{Q} = \mathrm{co}(V)$, $Q = \sum_{1\leq i\leq n} \lambda_i V_i$, where $\lambda_i \geq 0$ for all i and $\sum_i \lambda_i = 1$, so $A_1 Q = \sum_{1\leq i\leq n} \lambda_i A_1 V_i \in \mathrm{co}(A_1V)$. ■

References

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

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, **81** 609–666.

Angrist, J., Lang, D. and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, **1** 136–63.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, **90** 431–442.

Bai, Y., Huang, S., Moon, S., Shaikh, A. and Vytlacil, E. J. (2024). On the identifying power of generalized monotonicity for average treatment effects. Tech. rep., National Bureau of Economic Research.

Bai, Y., Huang, S. and Tabord-Meehan, M. (2025). Sharp Testable Implications of Encouragement Designs. ArXiv:2411.09808 [econ], URL http://arxiv.org/abs/2411.09808.

Bai, Y., Santos, A. and Shaikh, A. M. (2022). On testing systems of linear inequalities with known coefficients. Tech. rep.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92** 1171–1176.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, vol. 6.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2008). Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization. *American Economic Review*, **98** 351–356. URL https://www.aeaweb.org/articles?id=10.1257/aer.98.2.351.

Bhattacharya, J., Shaikh, A. M. and Vytlacil, E. (2012). Treatment effect bounds: An application to Swan–Ganz catheterization. *Journal of Econometrics*, **168** 223–243. URL https://www.sciencedirect.com/science/article/pii/S0304407612000024.

Bhuller, M. and Sigstad, H. (2024). 2sls with multiple treatments. *Journal of Econometrics*, **242** 105785. URL https://www.sciencedirect.com/science/article/pii/S0304407624001313.

Charnes, A. and Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, **9** 181–186. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800090303, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800090303.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68** 815–836.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, **162** 47–70.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, **82** 197–228.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, **86** 377–412.

Fang, Z., Santos, A., Shaikh, A. M. and Torgovitsky, A. (2023). Inference for large-scale linear systems with known coefficients. *Econometrica*, **91** 299–327.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58** 21–29.

Gabriel, E. E., Sachs, M. C. and Jensen, A. K. (2024). Sharp symbolic nonparametric bounds for measures of benefit in observational and imperfect randomized studies with ordinal outcomes. *Biometrika* asae020.

Goff, L. (2025). When does iv identification not restrict outcomes? 2406.02835, URL https://arxiv.org/abs/2406.02835.

Heckman, J. J. and Pinto, R. (2018). Unordered monotonicity. *Econometrica*, **86** 1–35.

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

Heckman, J. J. and Smith, J. A. (1998). Evaluating the welfare state. *ECONOMETRIC SOCIETY MONOGRAPHS*, **16** 241–318.

Huang, E. J., Fang, E. X., Hanley, D. F. and Rosenblum, M. (2019). Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data. *Biometrics*, **75** 1228–1239.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** 467–475.

Kirkeboen, L. J., Leuven, E. and Mogstad, M. (2016). Field of Study, Earnings, and Self-Selection. *The Quarterly Journal of Economics*, **131** 1057–1111. URL https://doi.org/10.1093/qje/qjw019.

Kitagawa, T. (2015). A Test for Instrument Validity. *Econometrica*, **83** 2043–2063. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11974, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11974.

Kline, P. and Santos, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics*, **4** 231–267.

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

Kwon, S. and Roth, J. (2024). Testing mechanisms. *arXiv preprint arXiv:2404.11739*.

Lee, S. and Salanié, B. (2023). Treatment Effects with Targeting Instruments. ArXiv:2007.10432 [econ, stat], URL http://arxiv.org/abs/2007.10432.

Lu, J., Ding, P. and Dasgupta, T. (2018). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, **43** 540–567.

Machado, C., Shaikh, A. M. and Vytlacil, E. J. (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics*.

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society* 1311–1334.

Manski, C. F. and Pepper, J. V. (1998). Monotone instrumental variables with an application to the returns to schooling.

Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via kmt constructions. *The Annals of probability* 266–291.

Masten, M. A. and Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, **11** 41–111.

Monrad, D. and Philipp, W. (1991). Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. *Probability Theory and Related Fields*, **88** 381–404. URL https://doi.org/10.1007/BF01418867.

Munkres, J. R. (2000). Topology. *Inc., Upper Saddle River*.

Navjeevan, M., Pinto, R. and Santos, A. (2023). Identification and estimation in a class of potential outcomes models. *arXiv preprint arXiv:2310.05311*.

Noack, C. (2021). Sensitivity of late estimates to violations of the monotonicity assumption. *arXiv preprint arXiv:2106.06421*.

Pollard, D. (2002). *A user's guide to measure theoretic probability*. 8, Cambridge University Press.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, **31** 567–584.

Sun, Z. (2023). Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, **237** 105523. URL https://www.sciencedirect.com/science/article/pii/S0304407623002397.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, **12** 389–434.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer-Verlag, New York. URL https://www.springer.com/gp/book/9780387946405.

Vytlacil, E. (2006). Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *The Review of Economics and Statistics*, **88** 578–581.

Online Appendix

E Auxiliary Lemmas

Lemma E.1. *Suppose Assumptions 4.1(i)-(iii) hold and let* $\Sigma(P) \equiv E_P[\psi(V,P)\psi(V,P)']$ *for* $\psi(V,P)$ *as defined in (26). If* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$*, then there exists* $\mathbb{G}(P) \sim N(0,\Sigma(P))$ *such that*

$$T_n = \sup_{s \in \mathcal{V}_n} \langle A^\dagger s, A^\dagger \mathbb{G}(P)\rangle + O_P\left(\frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) .$$

Proof. The claim follows from Lemma E.2 enabling us to apply identical arguments to those employed in Theorem 4.1 in Fang et al. (2023), but with their coupling result (their Lemma A.4) replaced by our improved coupling rate from Lemma E.3. ■

Lemma E.2. *If Assumptions 4.1(i)-(iii) hold and* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$ *then Assumptions 4.1, 4.2, 4.3, 4.4(v), and A.1 of Fang et al. (2023) are satisfied with* $a_n = \log(|\mathcal{M}|)\sqrt{|\mathcal{M}|/n}$ *and* $M_{3,\Psi} = \sqrt{|\mathcal{M}|}$.

Proof. Assumption 4.1(i) in Fang et al. (2023) is equivalent to our Assumption 4.1(i), while their Assumption 4.1(ii) holds with $a_n = \log(|\mathcal{M}|)\sqrt{|\mathcal{Z}|/n}$ by Lemma E.4 and Assumption 4.1(ii). Next note that Assumptions 4.2(i)(ii) in Fang et al. (2023) hold by Lemma E.8(i)(ii), their Assumption 4.2(iii) holds with $M_{3,\Psi} = \sqrt{|\mathcal{M}|}$ by Lemma E.8(iii), and their Assumption 4.2(iv) holds due to $\Omega(P)$ being diagonal. Further observe their Assumption 4.3(i) holds due the first $|\mathcal{M}|$ diagonal entries of $\Omega(P)$ being non-zero and the final two rows of $\psi(V,P)$ being zero. In turn, Assumptions 4.3(ii) and 4.4(v) in Fang et al. (2023) holds by Lemma E.8(iv) and Assumption 4.1(ii). Finally, we note that Assumption A.1 in Fang et al. (2023) holds with $a_n = \log(|\mathcal{M}|)\sqrt{|\mathcal{M}|/n}$ by Lemma E.6. ■

Lemma E.3. *Suppose Assumptions 4.1(i)(iii) hold,* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$*, and for* $\psi(V,P)$ *as in (26) let* $\Sigma(P) \equiv E_P[\psi(V,P)\psi(V,P)']$*. Then, (i) There exists a* $\mathbb{G}(P) \sim N(0,\Sigma(P))$ *satisfying uniformly in* $P \in \mathbf{P}$

$$\left\|\Omega(P)^\dagger\left(\sqrt{n}(\hat{\beta}_n - \beta(P)) - \mathbb{G}(P)\right)\right\|_\infty = O_P\left(\frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) . \tag{49}$$

(ii) There exists a $\mathbb{G}^\star(P) \sim N(0,\Sigma(P))$ *that is independent of* $\{V_i\}_{i=1}^n$ *and uniformly in*

$P \in \mathbf{P}$ *satisfies*

$$\left\|\Omega(P)^{\dagger}\left(\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n) - \mathbb{G}^{\star}(P)\right)\right\|_{\infty} = O_P\left(\left(\frac{\log^3(n)|\mathcal{M}|}{n}\right)^{1/4}\right) . \tag{50}$$

Proof. First note that Lemma E.4 allows us to conclude, uniformly in $P \in \mathbf{P}$, that we have

$$\|\Omega(P)^{\dagger}\{\sqrt{n}\{\hat{\beta}_n - \beta(P)\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(V_i, P)\}\|_{\infty} = O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{Z}|}}{\sqrt{n}}\right) . \tag{51}$$

To establish the lemma, we next apply Theorem 1 in Massart (1989) to couple the vector $\sum_i \psi(V_i, P)/\sqrt{n}$ to a Gaussian variable $\mathbb{G}_P$. To this end, we let $U_i, i = 1, \ldots, n$ be an i.i.d. sample with $U_i \sim \text{Unif}[0, 1]$. We also divide $[0, 1]$ into $|\mathcal{Z}|$ disjoint intervals, denoted $C_z(P)$, satisfying $P(U_i \in C_z(P)) = P_z$. We further subdivide each interval C_z into $|\mathcal{Y}||\mathcal{D}|$ disjoint subintervals, denoted $C_{ydz}(P)$, satisfying $P(U_i \in C_{ydz}(P)) = P_{ydz}$. In what follows, we will assume that (Y_i, D_i, Z_i) are generated from U_i according to the relation

$$I\{Y_i = y, D_i = d, Z_i = z\} = I\{U_i \in C_{ydz}(P)\} , \tag{52}$$

which we note is without loss of generality in that the distribution of (Y_i, D_i, Z_i) is still P in this probability space. Next, let $\mathcal{S}$ denote the class of all intervals contained in $[0, 1]$. Because all intervals are convex, $\mathcal{S}$ satisfies the uniform Minkowski condition in Definition 2 in Massart (1989). Moreover, since $\mathcal{S}$ is a VC-class, it follows that $\mathcal{S}$ satisfies assumption $H(\zeta)$ in Massart (1989) with $\zeta = 0$. Defining $\mathbb{G}_n \in \ell^{\infty}(\mathcal{S})$ by

$$\mathbb{G}_n(S) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(I\{U_i \in S\} - P(U \in S))$$

for any $S \in \mathcal{S}$, we then obtain from Theorem 1 in Massart (1989) that there is a Brownian bridge $\mathbb{W}$ satisfying

$$\text{Cov}[\mathbb{G}_n(S), \mathbb{G}_n(S')] = \text{Cov}[\mathbb{W}(S), \mathbb{W}(S')] \tag{53}$$

for all $S, S' \in \mathcal{S}$ and such that for some constants K, Λ, and θ depending only on $\mathcal{S}$ we have for all $t > 0$

$$P\left\{\sup_{S\in\mathcal{S}}|\mathbb{G}_n(S) - \mathbb{W}(S)| > \frac{\sqrt{\log(n)}}{\sqrt{n}}(t + K\log(n))\right\} \leq \Lambda\exp(-\theta t) . \tag{54}$$

In particular, we note that by setting t to be large enough in (54) we can conclude that we have

$$\sup_{S\in\mathcal{S}} |\mathbb{G}_n(S) - \mathbb{W}(S)| = O_P\left(\frac{\log^{3/2}(n)}{\sqrt{n}}\right) . \tag{55}$$

To conclude, define a Gaussian vector $\mathbb{G}(P)$ of dimension $|\mathcal{M}| + 2$ by letting its last two coordinates equal zero and its first $|\mathcal{M}|$ coordinates, indexed by $(y, d, z) \in \mathcal{M}$, be given by

$$\mathbb{G}_{ydz}(P) = \frac{1}{P_z}\mathbb{W}(C_{ydz}(P)) - \frac{P_{ydz}}{P_z^2}\mathbb{W}(C_z(P)) .$$

Similarly, note that by (52), the definition of $\psi(V, P)$, and the construction of $C_{ydz}(P)$ and $C_z(P)$ we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{ydz}(V_i, P) = \frac{1}{P_z}\mathbb{G}_n(C_{ydz}(P)) - \frac{P_{ydz}}{P_z^2}\mathbb{G}_n(C_z(P)) . \tag{56}$$

In particular, results (53) and (56) imply that $E[\mathbb{G}(P)\mathbb{G}(P)'] = \Sigma(P)$ as desired. Moreover, we have that

$$\begin{aligned}
&\left\|\Omega(P)^{\dagger}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(V_i, P) - \mathbb{G}(P)\right\}\right\|_{\infty} \\
&\qquad = \max_{(y,d,z)\in\mathcal{M}} \frac{P_z^{1/2}}{P_{yd|z}^{1/2}(1 - P_{yd|z})^{1/2}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{ydz}(V_i, P) - \mathbb{G}_{ydz}(P)\right| \\
&\lesssim \max_{(y,d,z)\in\mathcal{M}} \frac{P_z^{1/2}}{P_{yd|z}^{1/2}}\left(\frac{1}{P_z} + \frac{P_{ydz}}{P_z^2}\right) \times \sup_{S\in\mathcal{S}} |\mathbb{G}_n(S) - \mathbb{W}(S)| = O_P\left(\frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) ,
\end{aligned} \tag{57}$$

where the final result holds uniformly in $P \in \mathbf{P}$ by Assumption 4.1(iii) and result (55). The first claim of the lemma therefore follows from (51) and (57).

In order to establish the second claim of the lemma, we first define the vector $\hat{\mathbb{G}}_n$ to be given by

$$\hat{\mathbb{G}}_n \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi(V_i^*, P) - \frac{1}{n}\sum_{j=1}^{n}\psi(V_j, P)) \tag{58}$$

Given this notation, then note that Lemma E.5 allows us to conclude, uniformly in $P \in \mathbf{P}$, that

$$\|(\Omega(P))^{\dagger}\{\sqrt{n}\{\hat{\beta}_n^{\star} - \hat{\beta}_n\} - \hat{\mathbb{G}}_n\}\|_{\infty} = O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) . \tag{59}$$

Next, note that applying the same construction based on Massart (1989), but conditionally on $\{V_i\}_{i=1}^n$, implies that there is $\mathbb{G}(\hat{P})$ satisfying $\mathbb{G}(\hat{P}) \sim N(0, \mathrm{Var}_{\hat{P}}\{\psi(V, P)\})$ conditionally on $\{V_i\}_{i=1}^n$ and

$$P\left\{\|\hat{\mathbb{G}}_n - \mathbb{G}(\hat{P})\|_\infty > C_1\frac{|\mathcal{Z}|\sqrt{\log(n)}}{\sqrt{n}}(t + K\log(n))\Big|\{V_i\}_{i=1}^n\right\} \leq \Lambda\exp\{-\theta t\} \tag{60}$$

for some $C_1 < \infty$. Further let $\mathcal{B}$ denote the Borel σ-field, and for any $B \in \mathcal{B}$ denote its ϵ-enlargement under $\|\cdot\|_\infty$ by $B^\epsilon \equiv \{\tilde{b} : \inf_{b\in B}\|b-\tilde{b}\|_\infty \leq \epsilon\}$. Defining $\delta_n = 2C_1K\log^{3/2}(n)|\mathcal{Z}|/\sqrt{n}$ and setting $t = C_2$, we then obtain from result (60) and Strassen's Theorem (see, e.g., Theorem 10.3 in Pollard (2002)), that

$$\sup_{P\in\mathbf{P}} E_P\left[\sup_{B\in\mathcal{B}}\left\{P\left\{\hat{\mathbb{G}}_n \in B\Big|\{V_i\}_{i=1}^n\right\} - P\left\{\mathbb{G}(\hat{P}) \in A^{K\delta_n}\Big|\{V_i\}_{i=1}^n\right\}\right\}\right] \leq \Lambda\exp\{-\theta C_2\} \tag{61}$$

for n sufficiently large. Since the right hand side of (61) can be made arbitrarily small by setting C_2 sufficiently large, we obtain from Theorem 4 in Monrad and Philipp (1991) that there exists a random variable $\bar{\mathbb{G}}(P)$ with distribution $N(0, \text{Var}_{\hat{P}}\{\psi(V,P)\})$ conditionally on $\{V_i\}_{i=1}^n$ and such that uniformly in $P \in \mathbf{P}$

$$\|\hat{\mathbb{G}}_n - \bar{\mathbb{G}}(P)\|_\infty = O_P\left(\frac{\log^{3/2}(n)|\mathcal{Z}|}{\sqrt{n}}\right). \tag{62}$$

Moreover, result (62), the definition of $\Omega(P)$, and Assumption 4.1(iii) yield that uniformly in $P \in \mathbf{P}$

$$\begin{aligned}&\|(\Omega(P))^\dagger\{\hat{\mathbb{G}}_n - \bar{\mathbb{G}}(P)\}\|_\infty \\ &\qquad \leq \sup_{(y,d,z)\in\mathcal{M}}\left(\frac{P_z}{P_{yd|z}(1-P_{yd|z})}\right)^{1/2}\|\hat{\mathbb{G}}_n - \bar{\mathbb{G}}(P)\|_\infty = O_P\left(\frac{\log^{3/2}(n)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right).\end{aligned} \tag{63}$$

However, since $(\Omega(P))^\dagger\bar{\mathbb{G}}(P) \sim N(0, (\Omega(P))^\dagger\text{Var}_{\hat{P}}\{\psi(V,P)\}(\Omega(P))^\dagger)$ conditionally on the data, we may apply Lemma A.7 in Fang et al. (2023) together with Lemma E.7 to conclude that there exists a $\bar{\mathbb{G}}^\star(P) \sim N(0, (\Omega(P))^\dagger\text{Var}_P\{\psi(V,P)\}(\Omega(P))^\dagger)$ that is independent of the data, and in addition satisfies

$$\|(\Omega(P))^\dagger\bar{\mathbb{G}}(P) - \bar{\mathbb{G}}^\star(P)\|_\infty = O_P\left(\left(\frac{\log^3(|\mathcal{M}|)|\mathcal{M}|}{n}\right)^{1/4}\right) \tag{64}$$

uniformly in $P \in \mathbf{P}$. Finally, let $\mathbb{G}^\star(P) = \Omega(P)\bar{\mathbb{G}}^\star(P)$ and note that $\mathbb{G}^\star(P)$ is independent of the data, because $\bar{\mathbb{G}}^\star(P)$ is, and $\mathbb{G}^\star(P) \sim N(0, \text{Var}_P\{\psi(V,P)\})$ because the columns of $\text{Var}_P\{\psi(V,P)\}$ are in the range of $\Omega(P)$. The second claim of the lemma then follows from results (59), (63), and (64). ■

Lemma E.4. *Let Assumptions 4.1(i), 4.1(iii) hold and $\psi(V,P)$ be as defined in (26). If $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$, then it follows that, uniformly in $P \in \mathbf{P}$, we have*

$$\|(\Omega(P))^{\dagger}\{\sqrt{n}\{\hat{\beta}_n - \beta(P)\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(V_i,P)\}\|_{\infty} = O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{Z}|}}{\sqrt{n}}\right) .$$

Proof. We first note that by Lemma E.9(iv), $\min_{z\in\mathcal{Z}} \hat{P}_z > 0$ with probability tending to one uniformly in $P \in \mathbf{P}$. Therefore, the first $|\mathcal{M}|$ coordinates of $\hat{\beta}_n - \beta(P)$ have the following structure for some $(y,d,z) \in \mathcal{M}$

$$\frac{\hat{P}_{ydz}}{\hat{P}_z} - \frac{P_{ydz}}{P_z} = \frac{1}{P_z}(\hat{P}_{ydz} - P_{ydz}) - \frac{P_{ydz}}{P_z^2}(\hat{P}_z - P_z) + \hat{\gamma}_{ydz} ,$$

where

$$\hat{\gamma}_{ydz} = (\hat{P}_{ydz} - P_{ydz})\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right) - \frac{P_{ydz}}{P_z}\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)(\hat{P}_z - P_z) .$$

Moreover, since the final two rows of $\{\hat{\beta}_n - \beta(P)\}$ are identically zero, the definition of $\Omega(P)$ implies that

$$\begin{aligned}&\left\|(\Omega(P))^{\dagger}\left\{\sqrt{n}\{\hat{\beta}_n - \beta(P)\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(V_i,P)\right\}\right\|_{\infty} \\ &\quad = \max_{(y,d,z)\in\mathcal{M}}\left|\left(\frac{P_z}{P_{yd|z}(1-P_{yd|z})}\right)^{1/2}\sqrt{n}\hat{\gamma}_{ydz}\right| \lesssim \left(\frac{|\mathcal{Y}||\mathcal{D}|}{|\mathcal{Z}|}\right)^{1/2} \times \max_{(y,d,z)\in\mathcal{M}}\sqrt{n}\,|\hat{\gamma}_{ydz}| , \end{aligned} \tag{65}$$

where the second inequality follows from Assumption 4.1(iii). Next note Lemma E.9 allows us to conclude

$$\begin{aligned}&\max_{(y,d,z)\in\mathcal{M}}\left|(\hat{P}_{ydz} - P_{ydz})\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)\right| \\ &\leq \max_{(y,d,z)\in\mathcal{M}}|\hat{P}_{ydz} - P_{ydz}| \times \max_{z\in\mathcal{Z}}|\hat{P}_z - P_z| \times \max_{z\in\mathcal{Z}}\frac{1}{P_z^2} \times O_P(1) = O_P\left(\frac{\log(|\mathcal{M}|)|\mathcal{Z}|^2}{n\sqrt{|\mathcal{M}||\mathcal{Z}|}}\right)\end{aligned} \tag{66}$$

uniformly in $P \in \mathbf{P}$. Similarly, another application of Lemma E.9 and Assumption 4.1(iii) yield that

$$\begin{aligned}&\max_{(y,d,z)\in\mathcal{M}}\left|\frac{P_{ydz}}{P_z}\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)(\hat{P}_z - P_z)\right| \\ &\qquad \leq \max_{(y,d,z)\in\mathcal{M}}\frac{P_{ydz}}{P_z^3} \times \max_{z\in\mathcal{Z}}|\hat{P}_z - P_z|^2 \times O_P(1) = O_P\left(\frac{|\mathcal{Z}|^2\log(|\mathcal{Z}|)}{|\mathcal{M}|n}\right)\end{aligned} \tag{67}$$

uniformly in $P \in \mathbf{P}$. The claim of the lemma then follows from combining results (65), (66), and (67). ■

Lemma E.5. *Let Assumptions 4.1(i)(iii) hold and $\psi(V,P)$ be as defined in (26). If $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$, then it follows that, uniformly in $P \in \mathbf{P}$, we have*

$$\|(\Omega(P))^\dagger\{\sqrt{n}\{\hat{\beta}_n^\star - \hat{\beta}_n\} - \frac{1}{\sqrt{n}}\sum_{i=1}^n(\psi(V_i^*,P) - \frac{1}{n}\sum_{j=1}^n \psi(V_j,P))\}\|_\infty = O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) .$$

Proof. The proof follows similar arguments to those employed in the proof of Lemma E.4. We first note that by Lemmas E.9(iv) and E.10(iv), $\min_{z\in\mathcal{Z}} \hat{P}_z^* \wedge \hat{P}_z > 0$ with probability tending to one uniformly in $P \in \mathbf{P}$. Therefore, the first $|\mathcal{M}|$ coordinates of $\hat{\beta}_n^* - \hat{\beta}_n$ have the following structure for some $(y,d,z) \in \mathcal{M}$

$$\frac{\hat{P}_{ydz}^*}{\hat{P}_z^*} - \frac{\hat{P}_{ydz}}{\hat{P}_z} = \frac{1}{\hat{P}_z}(\hat{P}_{ydz}^* - \hat{P}_{ydz}) - \frac{\hat{P}_{ydz}}{\hat{P}_z^2}(\hat{P}_z^* - \hat{P}_z) + \hat{\gamma}_{ydz}^* , \tag{68}$$

where

$$\hat{\gamma}_{ydz}^* = (\hat{P}_{ydz}^* - \hat{P}_{ydz})\left(\frac{1}{\hat{P}_z^*} - \frac{1}{\hat{P}_z}\right) + \frac{\hat{P}_{ydz}}{\hat{P}_z}\left(\frac{1}{\hat{P}_z^*} - \frac{1}{\hat{P}_z}\right)(\hat{P}_z^* - \hat{P}_z) . \tag{69}$$

Moreover, since the final two rows of $\{\hat{\beta}_n - \hat{\beta}_n\}$ are identically zero, the definition of $\Omega(P)$ implies that

$$\begin{aligned}&\|(\Omega(P))^\dagger\{\sqrt{n}\{\hat{\beta}_n^* - \hat{\beta}_n\} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(V_i^*,\hat{P})\}\|_\infty \\ &\quad = \max_{(y,d,z)\in\mathcal{M}}\left|\left(\frac{P_z}{P_{yd|z}(1-P_{yd|z})}\right)^{1/2}\sqrt{n}\hat{\gamma}_{ydz}^*\right| \lesssim \left(\frac{|\mathcal{Y}||\mathcal{D}|}{|\mathcal{Z}|}\right)^{1/2} \times \max_{(y,d,z)\in\mathcal{M}}\sqrt{n}\left|\hat{\gamma}_{ydz}^*\right| , \end{aligned}\tag{70}$$

where the second inequality follows from Assumption 4.1(iii). Next note that Lemma E.10(iv) yields

$$\begin{aligned}&\max_{(y,d,z)\in\mathcal{M}}\left|(\hat{P}_{ydz}^* - \hat{P}_{ydz})\left(\frac{1}{\hat{P}_z^*} - \frac{1}{\hat{P}_z}\right)\right| \\ &\leq \max_{(y,d,z)\in\mathcal{M}}|\hat{P}_{ydz}^* - \hat{P}_{ydz}| \times \max_{z\in\mathcal{Z}}|\hat{P}_z^* - \hat{P}_z| \times \max_{z\in\mathcal{Z}}\frac{1}{\hat{P}_z^2} \times O_P(1) = O_P\left(\frac{\log(|\mathcal{M}|)|\mathcal{Z}|^2}{n\sqrt{|\mathcal{M}||\mathcal{Z}|}}\right)\end{aligned}\tag{71}$$

uniformly in $P \in \mathbf{P}$, where in the final result we used Lemmas E.10(i)(ii), Lemma E.9(iv), and Assumption 4.1(iii). Similarly, applying Lemmas E.10(iii)(iv) and Lemma E.9(iv) yield, uniformly in $P \in \mathbf{P}$, that

$$\max_{(y,d,z)\in\mathcal{M}} \left|\frac{\hat{P}_{ydz}}{\hat{P}_z}\left(\frac{1}{\hat{P}_z^*} - \frac{1}{\hat{P}_z}\right)(\hat{P}_z^* - \hat{P}_z)\right| \\ \leq \max_{(y,d,z)\in\mathcal{M}} \frac{P_{ydz}}{P_z^3} \times \max_{z\in\mathcal{Z}} |\hat{P}_z^* - \hat{P}_z|^2 \times O_P(1) = O_P\left(\frac{|\mathcal{Z}|^2\log(|\mathcal{Z}|)}{|\mathcal{M}|n}\right) \quad (72)$$

where in the final equality we used Lemma E.10(ii) and Assumption 4.1(iii). We therefore obtain that

$$\|(\Omega(P))^\dagger\{\sqrt{n}\{\hat{\beta}_n^* - \hat{\beta}_n\} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(V_i^*, \hat{P})\|_\infty = O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{Z}|}}{\sqrt{n}}\right) \quad (73)$$

uniformly in $P \in \mathbf{P}$, by combining results (70), (71), and (72). Next, note that for any $(y, d, z) \in \mathcal{M}$ we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{ydz}(V_i^*, \hat{P}) \\ = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\frac{1}{\hat{P}_z}(I\{Y_i^* = y, D_i^* = d, Z_i^* = z\} - \hat{P}_{ydz}) - \frac{\hat{P}_{ydz}}{\hat{P}_z^2}(I\{Z_i^* = z\} - \hat{P}_z)\right\}, \quad (74)$$

and therefore

$$\sum_{i=1}^n(\psi_{ydz}(V_i^*, \hat{P}) - (\psi_{ydz}(V_i^*, P) - \frac{1}{n}\sum_{j=1}^n \psi_{ydz}(V_j, P))) \\ = \sum_{i=1}^n \left\{\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)\left(I\{Y_i^* = y, D_i^* = d, Z_i^* = z\} - \hat{P}_{ydz}\right) \right. \\ \left. - \left(\frac{\hat{P}_{ydz}}{\hat{P}_z^2} - \frac{P_{ydz}}{P_z^2}\right)\left(I\{Z_i^* = z\} - \hat{P}_z\right)\right\}. \quad (75)$$

In particular, result (75), the definition of $\Omega(P)$, and Assumption 4.1(iii) allow us to conclude that

$$\|(\Omega(P))^\dagger\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(V_i^*, \hat{P}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n(\psi(V_i^*, P) - \frac{1}{n}\sum_{j=1}^n \psi(V_j, P))\}\|_\infty \\ \lesssim \max_{(y,d,z)\in\mathcal{M}} \left|\left(\frac{P_z}{P_{yd|z}}\right)^{1/2}\sqrt{n}\left\{\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)\left(\hat{P}_{ydz}^* - \hat{P}_{ydz}\right) - \left(\frac{\hat{P}_{ydz}}{\hat{P}_z^2} - \frac{P_{ydz}}{P_z^2}\right)\left(\hat{P}_z^* - \hat{P}_z\right)\right\}\right|. \quad (76)$$

Next, observe that Assumption 4.1(iii) and Lemmas E.9(ii)(iv) and E.10(i) imply, uniformly in $P \in \mathbf{P}$,

$$\max_{(y,d,z)\in\mathcal{M}} \left| \left(\frac{P_z}{P_{yd|z}}\right)^{1/2} \sqrt{n}\left(\frac{1}{\hat{P}_z} - \frac{1}{P_z}\right)\left(\hat{P}^*_{ydz} - \hat{P}_{ydz}\right) \right|$$

$$\lesssim \frac{\sqrt{n|\mathcal{M}|}}{|\mathcal{Z}|} \times \max_{z\in\mathcal{Z}} \frac{1}{\hat{P}_z P_z} \times \max_{z\in\mathcal{Z}} |\hat{P}_z - P_z| \times \max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}|$$

$$= O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{Z}|}}{\sqrt{n}}\right) . \tag{77}$$

Moreover, the triangle inequality together with Lemmas E.9(ii)(v) and Assumption 4.1(iii) imply that

$$\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{yd|z}P_z - P_{yd|z}\hat{P}_z|$$

$$\leq \max_{(y,d,z)\in\mathcal{M}} \left| |\hat{P}_{yd|z} - P_{yd|z}|P_z + P_{yd|z}|\hat{P}_z - P_z| \right| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{n|\mathcal{Z}|}}\right) \tag{78}$$

uniformly in $P \in \mathbf{P}$. Therefore, result (78), Assumption 4.1(iii), and Lemmas E.9(iv) and E.10(ii) yield

$$\max_{(y,d,z)\in\mathcal{M}} \left| \left(\frac{P_z}{P_{yd|z}}\right)^{1/2} \sqrt{n}\left(\frac{\hat{P}_{ydz}}{\hat{P}_z^2} - \frac{P_{ydz}}{P_z^2}\right)\left(\hat{P}^*_{z} - \hat{P}_{z}\right) \right|$$

$$\lesssim \frac{\sqrt{n|\mathcal{M}|}}{|\mathcal{Z}|} \times \max_{z\in\mathcal{Z}} \frac{1}{\hat{P}_z P_z} \times \max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{yd|z}P_z - P_{yd|z}\hat{P}_z| \times \max_{z\in\mathcal{Z}} |\hat{P}^*_z - \hat{P}_z|$$

$$= O_P\left(\frac{\log(|\mathcal{M}|)\sqrt{|\mathcal{M}|}}{\sqrt{n}}\right) \tag{79}$$

uniformly in $P \in \mathbf{P}$. The claim of the lemma then follows from results (73), (76), (77), and (79). ■

Lemma E.6. *Let Assumptions 4.1(i) and 4.1(iii) hold, and let* $\|A\|_{o,\infty} = \sup_{\|a\|_\infty \leq 1} \|Aa\|_\infty$ *for any* $p \times p$ *matrix* A*. Then,* $\|\Omega(P)^\dagger(\hat{\Omega}_n - \Omega(P))\|_{o,\infty} = O_P(\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|/n})$ *uniformly in* $P \in \mathbf{P}$*.*

Proof. First note that for any constant $c \in \mathbf{R}$, we have $|c-1| = |\sqrt{c}-1||\sqrt{c}+1|$ and therefore that $|\sqrt{c}-1| \leq |c-1|$. Therefore, employing the definitions of $\Omega(P)$ and $\hat{\Omega}_n$ yields that

$$\|\Omega(P)^{\dagger}(\hat{\Omega}_n-\Omega(P))\|_{o,\infty} = \max_{(y,d,z)\in\mathcal{M}}\left|\frac{\hat{P}_{yd|z}^{1/2}(1-\hat{P}_{yd|z})^{1/2}}{\hat{P}_z^{1/2}}\times\frac{P_z^{1/2}}{P_{yd|z}^{1/2}(1-P_{yd|z})^{1/2}}-1\right|$$
$$\le \max_{(y,d,z)\in\mathcal{M}}\left|\frac{\hat{P}_{ydz}(1-\hat{P}_{yd|z})}{\hat{P}_z^2}\times\frac{P_z^2}{P_{ydz}(1-P_{yd|z})}-1\right| . \tag{80}$$

Next, note that $(a^2-b^2)=(a-b)(a+b)$, Lemmas E.9(ii) and E.9(iv), and Assumption 4.1(ii) imply that

$$\max_{z\in\mathcal{Z}}\left|\frac{P_z^2}{\hat{P}_z^2}-1\right| \le \max_{z\in\mathcal{Z}}\frac{1}{\hat{P}_z^2}\times\max_{z\in\mathcal{Z}}|\hat{P}_z-P_z|\times\max_{z\in\mathcal{Z}}|\hat{P}_z+P_z| = O_P\left(\frac{\sqrt{\log(|\mathcal{Z}|)|\mathcal{Z}|}}{\sqrt{n}}\right) \tag{81}$$

uniformly in $P\in\mathbf{P}$. Moreover, similarly relying on Assumption 4.1(iii) and Lemma E.9(v) we can conclude

$$\max_{(y,d,z)\in\mathcal{M}}\left|\frac{(1-\hat{P}_{yd|z})}{(1-P_{yd|z})}-1\right| = \max_{(y,d,z)\in\mathcal{M}}\left|\frac{(P_{yd|z}-\hat{P}_{yd|z})}{(1-P_{yd|z})}\right| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{Z}|}}{\sqrt{n}}\right) \tag{82}$$

uniformly in $P\in\mathbf{P}$. The lemma then follows from (80), Lemma E.9(iii), and results (81) and (82). ■

Lemma E.7. *Let Assumptions 4.1(i)(iii) hold, $\psi(V,P)$ be as defined in (26) and for any $k\times k$ symmetric matrix M let $\|M\|_o$ denote its largest eigenvalue. If $\log(|\mathcal{M}|)|\mathcal{M}|/n=o(1)$, then uniformly in $P\in\mathbf{P}$*

$$\|(\Omega(P))^{\dagger}(\mathrm{Var}_{\hat{P}}\{\psi(V,P)\}-\mathrm{Var}_P\{\psi(V,P)\})(\Omega(P))^{\dagger}\|_o = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}}{\sqrt{n}}\right) \tag{83}$$

Proof. First define the vector $\bar{\psi}_n(P)\equiv\sum_i\psi(V_i,P)/n$ for notational simplicity. Next, note that for any unit-length vector v (under $\|\cdot\|_2$), the Cauchy-Schwarz inequality allows us to conclude

$$\|\bar{\psi}_n(P)\bar{\psi}_n(P)'v\|_2 = \|\bar{\psi}_n(P)\|_2|\bar{\psi}_n(P)'v| \le \|\bar{\psi}_n(P)\|_2^2 = \sum_{(y,d,z)\in\mathcal{M}}\left(\frac{1}{n}\sum_{i=1}^n\psi_{ydz}(V_i,P)\right)^2,$$

where the final equality follows from the definition of $\psi(V,P)$. Therefore, since $E_P[\psi(V,P)]=0$ for all $P\in\mathbf{P}$ and the sample $\{V_i\}_{i=1}^n$ is i.i.d., Assumption 4.1(iii) and the definition of $\Omega(P)$ yield the bound

$$\sup_{P\in\mathbf{P}} E_P\left[\|(\Omega(P))^{\dagger}\bar{\psi}_n\bar{\psi}_n'(\Omega(P))^{\dagger}\|_o\right] \leq \sup_{P\in\mathbf{P}}\left\{\|(\Omega(P))^{\dagger}\|_o^2 \times \frac{1}{n}\sum_{(y,d,z)\in\mathcal{M}} E_P\left[\psi_{ydz}^2(V,P)\right]\right\}$$

$$\leq \sup_{P\in\mathbf{P}}\left\{\max_{(y,d,z)\in\mathcal{M}} \frac{P_z}{P_{yd|z}(1-P_{yd|z})} \times \frac{1}{n}\sum_{(y,d,z)\in\mathcal{M}}\left(\frac{P_{ydz}}{P_z^2}+\frac{P_{ydz}^2}{P_z^3}\right)\right\} \lesssim \frac{|\mathcal{M}|}{n} . \quad (84)$$

Next, define the matrices $M_i(P) \equiv (\Omega(P))^{\dagger}(\psi(V_i,P)\psi(V_i,P)' - E_P[\psi(V,P)\psi(V,P)'])(\Omega(P))^{\dagger}/n$, and note that result (84), the triangle inequality, and Markov's inequality yield, uniformly in $P \in \mathbf{P}$, that

$$\|(\Omega(P))^{\dagger}\{\text{Var}_{\hat{P}}\{\psi(V,P)\} - \text{Var}_P\{\psi(V,P)\}\}(\Omega(P))^{\dagger}\|_o \leq \|\sum_{i=1}^n M_i(P)\|_o + O_P\left(\frac{|\mathcal{M}|}{n}\right) . \quad (85)$$

Further note that for any unit length vector v (under $\|\cdot\|_2$) we obtain by the Cauchy-Schwarz inequality

$$\|\psi(V,P)\psi(V,P)'v\|_2 \leq \|\psi(V,P)\|_2^2$$

$$= \sum_{(y,d,z)\in\mathcal{M}}\left(\frac{1}{P_z}I\{Y=y,D=d,Z=z\} - \frac{P_{ydz}}{P_z^2}I\{Z=z\}\right)^2$$

$$\leq 2\max_{z\in\mathcal{Z}} \sum_{(y,d)\in\mathcal{Y}\times\mathcal{D}}\left(\frac{1}{P_z^2}I\{Y=y,D=d\} + \frac{P_{ydz}^2}{P_z^4}\right)$$

$$\leq 2\max_{z\in\mathcal{Z}}\frac{1}{P_z^2} + 2\max_{z\in\mathcal{Z}}\sum_{(y,d)\in\mathcal{Y}\times\mathcal{D}}\frac{P_{ydz}^2}{P_z^4} \lesssim |\mathcal{Z}|^2 , \quad (86)$$

where the first equality follows by definition of $\psi(V,P)$ and the final inequality from Assumption 4.1(iii) and

$$\sum_{(y,d)\in\mathcal{Y}\times\mathcal{D}}\frac{P_{ydz}^2}{P_z^4} \leq \sum_{(y,d)\in\mathcal{Y}\times\mathcal{D}}\frac{P_{ydz}}{P_z^3} = \frac{1}{P_z^2} .$$

Hence, the definitions of $\Omega(P)$ and $M_i(P)$, the triangle inequality, result (86), and Lemma E.8(ii) imply

$$\|M_i(P)\|_o \leq \frac{1}{n}\{\|(\Omega(P))^{\dagger}\|_o^2\|\psi(V,P)\psi(V,P)'\|_o + \|(\Omega(P))^{\dagger}E_P[\psi(V,P)\psi(V,P)'](\Omega(P))^{\dagger}\|_o\}$$

$$\lesssim \frac{1}{n}\left(\sup_{(y,d,z)\in\mathcal{M}}\frac{P_z}{P_{yd|z}(1-P_{yd|z})}\times|\mathcal{Z}|^2 + O(1)\right) = \frac{|\mathcal{M}|}{n} + O\left(\frac{1}{n}\right) . \quad (87)$$

Next, observe that the definition of $\Omega(P)$, result (86), and Assumption 4.1(iii) allow us to conclude that

$$\|(\Omega(P))^{\dagger}\psi(V,P)\psi(V,P)'(\Omega(P))^{\dagger}\|_o \leq \max_{(y,d,z)\in\mathcal{M}} \frac{P_z}{P_{yd|z}(1-P_{yd|z})} \times \|\psi(V,P)\|_2^2 \lesssim |\mathcal{M}| \ . \quad (88)$$

Therefore, the triangle inequality, $\{M_i(P)\}_{i=1}^n$ being i.i.d., Lemma E.8(ii), and result (88) imply that

$$\|\sum_{i=1}^n E_P[M_i^2(P)]\|_o \leq \frac{1}{n}\left\{\|E_P[((\Omega(P))^{\dagger}\psi(V,P)\psi(V,P)'(\Omega(P))^{\dagger})^2]\|_o + O(1)\right\} \lesssim \frac{|\mathcal{M}|}{n} \ . \quad (89)$$

Together, results (87) and (89) allow us to apply Bernstein's inequality for matrices (see, e.g., Theorem 1.4 in Tropp, 2012) to conclude that there is a constant $C<\infty$ such that for all $t\geq 0$ we have

$$P\left\{\|\sum_{i=1}^n M_i(P)\|_o > t\right\} \leq |\mathcal{M}|\exp\left\{-C\frac{nt^2}{|\mathcal{M}|(1+t)}\right\} \ . \quad (90)$$

Hence, evaluating the bound in (90) at $t = K\sqrt{|\mathcal{M}|\log(|\mathcal{M}|)}/\sqrt{n}$ for K sufficiently large implies that

$$\|\sum_{i=1}^n M_i(P)\|_o = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}}{\sqrt{n}}\right) \quad (91)$$

uniformly in $P\in\mathbf{P}$. The claim of the lemma then follows from (85) and (91). ∎

Lemma E.8. *Let Assumptions 4.1(i), 4.1(ii) hold,* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$, $\psi(V,P)$ *be as in* (26)*, and* $\Sigma(P)\equiv E_P[\psi(V,P)\psi(V,P)']$*. Then, (i)* $E_P[\psi(V,P)]=0$*; (ii) The eigenvalues of* $(\Omega(P))^{\dagger}\Sigma(P)\Omega(P)^{\dagger}$ *are uniformly bounded in* $P\in\mathbf{P}$*; (iii)* $\Psi(V,P)\equiv\|\Omega(P)^{\dagger}\psi(V,P)\|_{\infty}$ *satisfies* $\sup_{P\in\mathbf{P}} E_P[|\Psi(V,P)|^3]\lesssim|\mathcal{M}|^{3/2}$*; and (iv)* $(\hat{\beta}_n-\beta(P)),(\hat{\beta}_n^*-\hat{\beta}_n)\in\mathrm{range}(\Sigma(P))$ *with probability tending to one uniformly in* $P\in\mathbf{P}$.

Proof. In order to establish the first claim, recall that the first $|\mathcal{M}|$ entries of $\psi(V,P)$ have the structure

$$\psi_{ydz}(V,P) = \frac{1}{P_z}I\{Y=y,D=d,Z=z\} - \frac{P_{ydz}}{P_z^2}I\{Z=z\} \ .$$

By direct calculation, $E_P[\psi_{ydz}(V,P)] = P_{yd|z}-P_{yd|z}=0$ and, since the final two coordinates of $\psi(V,P)$ are identically equal to zero, the first claim of the lemma follows.

To establish the second claim, we first characterize the covariance matrix $\Sigma(P)$ by noting that

$$\text{Cov}[\psi_{ydz}(V,P),\psi_{\tilde{y}\tilde{d}\tilde{z}}(V,P)] = \begin{cases} 0 & \text{if } z \neq \tilde{z} \\ -\frac{P_{yd|z}P_{\tilde{y}\tilde{d}|z}}{P_z} & \text{if } z = \tilde{z} \text{ and } (y,d) \neq (\tilde{y},\tilde{z}) \\ \frac{P_{yd|z}(1-P_{yd|z})}{P_z} & \text{if } (y,d,z) = (\tilde{y},\tilde{d},\tilde{z}) \end{cases} . \tag{92}$$

Also observe that the last two rows and columns of $\Sigma(P)$ are zero because the final two rows of $\psi(V,P)$ equal zero. Since $\Omega(P)$ is a diagonal matrix in which the first $|\mathcal{M}|$ entries have the structure $(P_{ydz}(1-P_{ydz})/P_z)^{1/2}$ for some $(y,d,z) \in \mathcal{M}$ and the final two entries equal to zero, result (92) implies that $\Omega(P)^\dagger\Sigma(P)\Omega(P)^\dagger$ is block diagonal and the final two columns and rows equal zero. In particular, there are $|\mathcal{Z}|$ blocks, which we denote by $\Gamma_z(P)$. Each $\Gamma_z(P)$ is a $|\mathcal{Y}||\mathcal{D}| \times |\mathcal{Y}||\mathcal{D}|$ matrix, with off-diagonal elements given by

$$-\frac{P_{yd|z}^{1/2}P_{\tilde{y}\tilde{d}|z}^{1/2}}{(1-P_{yd|z})^{1/2}(1-P_{\tilde{y}\tilde{d}|z})^{1/2}}$$

for some $(y,d) \neq (\tilde{y},\tilde{d})$, and diagonal elements equal to one. Next, note that Assumption 4.1(iii) gives us

$$\begin{aligned} \sup_{\|a\|_2=1} \|\Gamma_z(P)\|_2^2 &= \sup_{\|a\|_2=1} \sum_{(y,d)} \left(a_{yd} - \sum_{(\tilde{y},\tilde{d})\neq(y,d)} a_{\tilde{y}\tilde{d}} \times \frac{P_{yd|z}^{1/2}P_{\tilde{y}\tilde{d}|z}^{1/2}}{(1-P_{yd|z})^{1/2}(1-P_{\tilde{y}\tilde{d}|z})^{1/2}} \right)^2 \\ &\lesssim \sup_{\|a\|_2=1} \sum_{(y,d)} \left(|a_{yd}| + P_{yd|z}^{1/2} \sum_{(\tilde{y},\tilde{d})\neq(y,d)} |a_{\tilde{y}\tilde{d}}| P_{\tilde{y}\tilde{d}|z}^{1/2} \right)^2 \\ &\leq \sup_{\|a\|_2=1} 2 \sum_{(y,d)} \left(a_{yd}^2 + P_{yd|z} \sum_{(\tilde{y},\tilde{d})\neq(y,d)} a_{\tilde{y}\tilde{d}}^2 \right) \leq 4 , \end{aligned} \tag{93}$$

where the second inequality follows from the Cauchy-Schwarz inequality and $(b+c)^2 \leq 2(b^2+c^2)$ for any constants b, c, and the final inequality from $\sum_{(y,d)}(a_{yd}^2 + P_{yd|z}) = 2$ due to $\|a\|_2^2 = 1$. Since $\Omega(P)^\dagger\Sigma(P)\Omega(P)^\dagger$ is block diagonal with blocks $\Gamma_z(P)$, the second claim of the Lemma follows from (93).

For the third claim of the lemma, simply employ the defintion of $\psi(V,P)$ and $\Omega(P)$ to obtain that

$$\|\Omega(P)^\dagger\psi(V,P)\|_\infty$$

$$= \max_{(y,d,z)\in\mathcal{M}} \left(\frac{\sqrt{P_z}}{\sqrt{P_{yd|z}(1-P_{yd|z})}} \times \left| \frac{1}{P_z} I\{Y=y, D=d, Z=z\} - \frac{P_{ydz}}{P_z^2} I\{Z=z\} \right| \right)$$
$$\leq \max_{(y,d,z)\in\mathcal{M}} \frac{\sqrt{P_z}}{\sqrt{P_{yd|z}(1-P_{yd|z})}} \times \max_{(y,d,z)\in\mathcal{M}} \left| \frac{1}{P_z} + \frac{P_{yd|z}}{P_z} \right| \lesssim \sqrt{|\mathcal{M}|} \,,$$

where the final inequality holds uniformly in $P \in \mathbf{P}$ by Assumption 4.1(iii).

Turning to the fourth claim, we first note that by Lemma E.9(iii), the event $E_0 \equiv \{\min_{z\in\mathcal{Z}} \hat{P}_z > 0\}$ has probability tending to one uniformly in $P \in \mathbf{P}$. We next argue that $(\hat{\beta}_n - \beta(P)) \in \text{range}(\Sigma(P))$ whenever the event E_0 occurs. To this end, note that by result (92), $\Sigma(P)$ is block diagonal with $|\mathcal{Z}|$ blocks we denote by Λ_z, and has the final two rows and columns equal to zero. To examine the null space of Λ_z note that a vector $a \equiv (a_{yd} : (y,d) \in \mathcal{Y} \times \mathcal{D})$ satifies $\Lambda_z a = 0$ if and only if it satisfies the equality

$$0 = \text{Var}_P \left(\sum_{(y,d)} a_{yd} \left(\frac{1}{P_z} I\{Y=y, D=d, Z=z\} - \frac{P_{ydz}}{P_z^2} I\{Z=z\} \right) \right) \,. \tag{94}$$

Defining a function f of (Y,D) by $f(y,d) = a_{yd}$, condition (94) may be equivalently be expressed as f satisfying $(f(Y,D) - E_P[f(Y,D)|Z=z])I\{Z=z\} = 0$. However, by Assumption 4.1(iii) such an equality can only hold if f, and hence a, is constant. Given the structure of $\Sigma(P)$, we thus obtain that the null space of Σ has dimension $|\mathcal{Z}|+2$ and its basis is given by: (a) $|\mathcal{Z}|$ vectors whose final two coordinates equal zero and first $|\mathcal{M}|$ coordinates equal a vector $a \equiv (a_{ydz} : (y,d,z) \in \mathcal{M})$ satisfying $a_{ydz} = I\{z = z_0\}$ for some z_0; and (b) The coordinate vectors for the last two coordinates. Because the final two coordinates of $(\hat{\beta}_n - \beta(P))$ equal zero and in addition $\sum_{(y,d)} \hat{P}_{yd|z} = \sum_{(y,d)} P_{yd|z} = 1$ whenever the even E_0 occurs, it follows that $(\hat{\beta}_n - \beta(P))$ is orthogonal to the null space of $\Sigma(P)$ and hence belongs to its range because $\Sigma(P)$ is symmetric. Identical arguments further imply that $(\hat{\beta}_n^* - \hat{\beta}_n) \in \text{range}(\Sigma(P))$ with probability tending to one uniformly in $P \in \mathbf{P}$. ■

Lemma E.9. *Let Assumptions 4.1(i) and 4.1(iii) be satisfied. If in addition* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$*, then it follows that uniformly in* $P \in \mathbf{P}$*:*

(i) $\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz} - P_{ydz}| = O_P(\sqrt{\log(|\mathcal{M}|)}/\sqrt{|\mathcal{M}|n})$*;*

(ii) $\max_{z\in\mathcal{Z}} |\hat{P}_z - P_z| = O_P(\sqrt{\log(|\mathcal{Z}|)}/\sqrt{|\mathcal{Z}|n})$*;*

(iii) $\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz}/P_{ydz} - 1| = O_P(\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}/\sqrt{n})$*;*

(iv) $\max_{z\in\mathcal{Z}} |\hat{P}_z/P_z - 1| = O_P(\sqrt{\log(|\mathcal{Z}|)|\mathcal{Z}|}/\sqrt{n})$*;*

(v) $\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{yd|z} - P_{yd|z}| = O_P\sqrt{\log(|\mathcal{M}|)|\mathcal{Z}|/n})$*.*

PROOF. First note Bernstein's inequality (see, e.g., Lemma 2.2.9 in van der Vaart and Wellner, 1996) implies

$$P\left\{|\hat{P}_{ydz} - P_{ydz}| > x\right\} \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{\frac{P_{ydz}(1-P_{ydz})}{n} + \frac{x}{3n}}\right) ,$$

where the inequality holds for all $(y,d,z) \in \mathcal{M}$ and $P \in \mathbf{P}$. Therefore, Lemma 2.2.10 in van der Vaart and Wellner (1996) and the norm inequality $\|\cdot\|_1 \leq \|\cdot\|_{\psi_1}$ allow us to conclude that

$$E\left[\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz} - P_{ydz}|\right] \lesssim \frac{\log(|\mathcal{M}|)}{n} + \max_{(y,d,z)\in\mathcal{M}} (P_{ydz}(1-P_{ydz}))^{1/2}\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{n}} . \quad (95)$$

Next observe that $P_{ydz}(1-P_{ydz}) \lesssim 1/|\mathcal{M}|$ uniformly in $(y,d,z) \in \mathcal{M}$ and $P \in \mathbf{P}$ by Assumption 4.1(iii). Therefore, result (95) together with Markov's inequality and $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$ imply

$$\max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz} - P_{ydz}| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{|\mathcal{M}|n}}\right) \quad (96)$$

uniformly in $P \in \mathbf{P}$, which establishes the first claim of the lemma. Moreover, by identical arguments

$$\max_{z\in\mathcal{Z}} |\hat{P}_z - P_z| = O_P\left(\frac{\sqrt{\log(|\mathcal{Z}|)}}{\sqrt{|\mathcal{Z}|n}}\right) \quad (97)$$

uniformly in $P \in \mathbf{P}$, which establishes the second claim of the lemma. For the third claim, note that

$$\max_{(y,d,z)\in\mathcal{M}} \left|\frac{\hat{P}_{ydz}}{P_{ydz}} - 1\right| \leq \max_{(y,d,z)\in\mathcal{M}} \frac{1}{P_{ydz}} \times \max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz} - P_{ydz}| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}}{\sqrt{n}}\right) , \quad (98)$$

where the equality holds uniformly in $P \in \mathbf{P}$ by Assumption 4.1(iii) and result (96). The fourth claim of the lemma follows by identical arguments to those employed in (98) by relying on (97) instead of (96). To establish the final claim of the Lemma, we rely on the triangle inequality to obtain that

$$\begin{aligned}\max_{(y,d,z)\in\mathcal{M}} \left|\frac{\hat{P}_{ydz}}{\hat{P}_z} - \frac{P_{ydz}}{P_z}\right| &\leq \max_{z\in\mathcal{Z}} \frac{1}{\hat{P}_z} \times \max_{(y,d,z)\in\mathcal{M}} |\hat{P}_{ydz} - P_{ydz}| + \max_{(y,d,z)\in\mathcal{M}} P_{ydz} \times \max_{z\in\mathcal{Z}} \frac{|\hat{P}_z - P_z|}{\hat{P}_z P_z} \\ &= O_P\left(|\mathcal{Z}|\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{|\mathcal{M}|n}}\right) + O_P\left(\frac{|\mathcal{Z}|^2\sqrt{\log(|\mathcal{Z}|)}}{|\mathcal{M}|\sqrt{|\mathcal{Z}|n}}\right) = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{Z}|}}{\sqrt{n}}\right) \quad (99)\end{aligned}$$

where the equalities holds uniformly in $P \in \mathbf{P}$ by parts (i), (ii), and (iii) of this Lemma and

Assumption 4.1(iii). ■

Lemma E.10. *Let Assumptions 4.1(i) and 4.1(iii) be satisfied. If in addition* $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$*, then it follows that uniformly in* $P \in \mathbf{P}$*:*

(i) $\max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| = O_P(\sqrt{\log(|\mathcal{M}|)}/\sqrt{|\mathcal{M}|n})$*;*

(ii) $\max_{z\in\mathcal{Z}} |\hat{P}^*_z - \hat{P}_z| = O_P(\sqrt{\log(|\mathcal{Z}|)}/\sqrt{|\mathcal{Z}|n})$*;*

(iii) $\max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz}/\hat{P}_{ydz} - 1| = O_P(\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}/\sqrt{n})$*;*

(iv) $\max_{z\in\mathcal{Z}} |\hat{P}^*_z/\hat{P}_z - 1| = O_P(\sqrt{\log(|\mathcal{Z}|)|\mathcal{Z}|}/\sqrt{n})$.

Proof. We follow similar arguments to those used in the proof of Lemma E.9. First define the event

$$A_n \equiv \left\{ \max_{(y,d,z)\in\mathcal{M}} \left| \frac{\hat{P}_{ydz}}{P_{ydz}} - 1 \right| > \frac{1}{2} \text{ and } \max_{z\in\mathcal{Z}} \left| \frac{\hat{P}_z}{P_z} - 1 \right| > \frac{1}{2} \right\} ,$$

which we note is a function of $\{V_i\}_{i=1}^n$. Therefore, using the law of iterated expectations we obtain the bound

$$P\left\{ \max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| > x \right\} \leq E_P\left[P\left\{ \max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| > x \Big| \{V_i\}_{i=1}^n \right\} I\{A_n\} \right] + P\{A_n^c\} \quad (100)$$

for any $x > 0$. Next, note Bernstein's inequality (see Lemma 2.2.9 in van der Vaart and Wellner (1996)) implies

$$P\left\{ |\hat{P}^*_{ydz} - \hat{P}_{ydz}| > x \Big| \{V_i\}_{i=1}^n \right\} \leq 2\exp\left(-\frac{1}{2} \frac{x^2}{\frac{\hat{P}_{ydz}(1-\hat{P}_{ydz})}{n} + \frac{x}{3n}} \right) ,$$

where the inequality holds for all $(y,d,z) \in \mathcal{M}$. Therefore, Lemma 2.2.10 in van der Vaart and Wellner (1996) and the norm inequality $\|\cdot\|_1 \leq \|\cdot\|_{\psi_1}$ allow us to conclude that

$$E_P\left[\max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| \Big| \{V_i\}_{i=1}^n \right] I\{A_n\} \lesssim \frac{\log(|\mathcal{M}|)}{n} + \max_{(y,d,z)\in\mathcal{M}} P^{1/2}_{ydz} \frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{n}} . \quad (101)$$

Next observe that $P_{ydz} \lesssim 1/|\mathcal{M}|$ uniformly in $(y,d,z) \in \mathcal{M}$ and $P \in \mathbf{P}$ by Assumption 4.1(iii). Therefore, result (101) together with Markov's inequality and $\log(|\mathcal{M}|)|\mathcal{M}|/n = o(1)$ yield for any $K > 0$ that

$$E_P\left[P\left\{ \max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| > K\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{|\mathcal{M}|n}} \Big| \{V_i\}_{i=1}^n \right\} I\{A_n\} \right] \lesssim \frac{1}{K} . \quad (102)$$

Hence, since $\sup_{P\in\mathbf{P}} P\{A_n^c\} = o(1)$ by Lemmas E.9(iii)(iv), results (100) and (102) together imply that

$$\max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)}}{\sqrt{|\mathcal{M}|n}}\right) \tag{103}$$

uniformly in $P \in \mathbf{P}$, which establishes the first claim of the lemma. Moreover, by identical arguments,

$$\max_{z\in\mathcal{Z}} |\hat{P}^*_z - \hat{P}_z| = O_P\left(\frac{\sqrt{\log(|\mathcal{Z}|)}}{\sqrt{|\mathcal{Z}|n}}\right) \tag{104}$$

uniformly in $P \in \mathbf{P}$, which establishes the second claim of the lemma. For the third claim, note that

$$\max_{(y,d,z)\in\mathcal{M}} \left|\frac{\hat{P}^*_{ydz}}{\hat{P}_{ydz}} - 1\right| \leq \max_{(y,d,z)\in\mathcal{M}} \frac{1}{\hat{P}_{ydz}} \times \max_{(y,d,z)\in\mathcal{M}} |\hat{P}^*_{ydz} - \hat{P}_{ydz}| = O_P\left(\frac{\sqrt{\log(|\mathcal{M}|)|\mathcal{M}|}}{\sqrt{n}}\right) , \tag{105}$$

where the equality holds uniformly in $P \in \mathbf{P}$ by Lemma E.9(iii), Assumption 4.1(iii), and result (103). The fourth claim of the lemma follows by identical arguments to those employed in (105) but relying on (104) instead of (103) and Lemma E.9(iv) instead of Lemma E.9(iii). ■

Lemma E.11. *Let Assumption 5.1 hold,* $\operatorname{diam}\{B_{\ell,L}\} < \delta$ *for all* $1 \leq \ell \leq L$*, a function* $f : \mathcal{M} \to \mathbf{R}$ *satisfy*

$$\max_{d\in\mathcal{D},z\in\mathcal{Z}} \sup_{|y-y'|\leq 2\delta} |f(y,d,z) - f(y',d,z)| \leq \eta \tag{106}$$

for some $\eta > 0$*, and define* $\mathcal{I}(\mathcal{R}) \equiv \{(c_o, r_t, \chi) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t : (\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}) \cap \mathcal{R} \neq \emptyset\}$*. Further suppose* $P = QT^{-1}$ *where* $Q = Q_R \times P_Z$ *for* P_Z *the marginal distribution of* Z *under* P *and* Q_R *a distribution for* R *satisfying the restrictions*

$$\sum_{(c_o,r_t,\chi)\in\mathcal{I}(\mathcal{R})} Q_R\{(R_o, R_t) \in (\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\})\} = 1 \tag{107}$$

$$\sum_{(c_o,r_t)\in\mathcal{L}^{|\mathcal{D}|}\times\mathcal{R}_t} \left(\sup_{r_o\in B_{c_o}\cap\mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0\right) \times Q'_R\{(R_o, R_t) \in (B_{c_o} \times \{r_t\}) \cap \mathcal{R}'\} \geq 0 \tag{108}$$

$$\sum_{(c_o,r_t)\in\mathcal{L}^{|\mathcal{D}|}\times\mathcal{R}_t} \left(\inf_{r_o\in B_{c_o}\cap\mathcal{R}'_o(r_t)} g(r_o, r_t) - \theta_0\right) \times Q'_R\{(R_o, R_t) \in (B_{c_o} \times \{r_t\}) \cap \mathcal{R}'\} \leq 0 \ . \tag{109}$$

Then, there is a distribution $\tilde{Q}_R$ *of* R *satisfying* $\tilde{Q}_R\{R \in \mathcal{R}\} = 1$*,* $E_{\tilde{Q}}[(g(R) - \theta_0)I\{R \in \mathcal{R}'\}] = 0$*, and such that* $\tilde{P} = \tilde{Q}T^{-1}$ *with* $\tilde{Q} = \tilde{Q}_R \times P_Z$ *satisfies* $\int f(dP - d\tilde{P}) \leq 2\eta$*.*

Proof. The proof proceeds by constructing a discrete measure $\tilde{Q}_R$ satisfying the desired properties. We distinguish multiple cases depending on whether inequalities (108) and (109) hold strictly or with equality. In what follows, it will be helpful to note that $P = QT^{-1}$ and $R \perp\!\!\!\perp Z$ under Q imply that

$$P\{Y \in B_{\ell,L}, D = d, Z = z\} = \sum_{c_o \in \mathcal{L}^{|\mathcal{D}|}, r_t \in \mathcal{R}_t} Q\{R_o \in B_{c_o}, R_t = r_t\} I\{c_o(d) = \ell, r_t(z) = d\} P\{Z = z\} \ . \quad (110)$$

Moreover, the same equality holds if we replace P and Q with $\tilde{P}$ and $\tilde{Q}$. In addition, note that if $(c_o, r_t, \chi) \in I(\mathcal{R})$, then $\mathcal{Y}(c_o, r_t, \chi) \neq \emptyset$.

Case I. First, suppose that the inequalities in (108) and (109) both hold strictly. To address this case, define

$$\mathbf{V} \equiv \bigotimes_{(c_o, r_t) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t : B_{c_o} \cap \mathcal{R}'_o(r_t) \neq \emptyset} (B_{c_o} \cap \mathcal{R}'_o(r_t)) \ .$$

Next note $\mathbf{V}$ is connected because $B_{c_o} \cap \mathcal{R}'_o(r_t)$ is connected for each $c_o \in \mathcal{L}^{|\mathcal{D}|}$ and $r_t \in \mathcal{R}_t$ by Assumption 5.1(iv). Write an element $v \in \mathbf{V}$ as $v = (v(c_o, r_t) \in B_{c_o} \cap \mathcal{R}'_o(r_t))$ and define

$$F(v) \equiv \sum_{(c_o, r_t) \in L^{|\mathcal{D}|} \times \mathcal{R}_t : B_{c_o} \cap \mathcal{R}'_o(r_t) \neq \emptyset} (g(v(c_o, r_t), r_t) - \theta_0) \times Q\{(R_o, R_t) \in (B_{c_o} \times \{r_t\}) \cap \mathcal{R}'\} \ .$$

Note that $F : \mathbf{V} \to \mathbf{R}$ and that inequalities (108) and (109) holding strictly imply there are $\underline{v}, \bar{v} \in \mathbf{V}$ satisfying $F(\underline{v}) < 0 < F(\bar{v})$. Since $F : \mathbf{V} \to \mathbf{R}$ is continuous and $\mathbf{V}$ is connected, it follows that $F(\mathbf{V})$ is connected as well; see, e.g., Theorem 23.5 in Munkres (2000). Therefore, we conclude that there is a $v^\star \in \mathbf{V}$ such that $F(v^\star) = 0$. Finally, define a discrete measure $\tilde{Q}_R$ to have support points $(s(c_o, r_t, \chi) : (c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R}))$, where

$$s(c_o, r_t, \chi) = \begin{cases} (v^\star(c_o, r_t), r_t) & \text{if } \chi = 1 \\ \text{any } (r_o, r_t) \in (\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}) \cap \mathcal{R} & \text{if } \chi = 0 \ . \end{cases}$$

We also assign probabilities $\tilde{Q}_R\{(R_o, R_t) = s(c_o, r_t, \chi)\} = Q_R\{(R_o, R_t) \in \mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}\}$ for any $(c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R})$. Then note that by (106) we have $\tilde{Q}\{R \in \mathcal{R}\} = 1$, while $F(v^\star) = 0$ implies $E_{\tilde{Q}_R}[(g(R) - \theta_0) I\{R \in \mathcal{R}'\}] = 0$. Next, use result (110) together with Q_R satisfying restriction (107) to obtain the upper bound

$$\int f dP \leq \sum_{(\ell, d, z) \in \mathcal{M}_L} P\{Y \in B_{\ell,L}, D = d, Z = z\} \times \sup_{y \in B_{\ell,L}} f(y, d, z)$$
$$= \sum_{(\ell, d, z) \in \mathcal{M}_L} \sum_{(c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R})} Q\{R_o \in \mathcal{Y}(c_o, r_t, \chi), R_t = r_t\} I\{c_o(d) = \ell, r_t(z) = d\} P\{Z = z\}$$

$$\times \sup_{y \in B_{\ell,L}} f(y,d,z) \ . \quad (111)$$

Similarly, $\tilde{Q}_R\{(R_o, R_t) = s(c_o, r_t, \chi)\} = Q_R\{(R_o, R_t) \in \mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}\}$ for $(c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R})$ and $\tilde{P}_Z = P_Z$ imply

$$\int f d\tilde{P} \geq \sum_{(\ell,d,z)\in\mathcal{M}_L} \tilde{P}\{Y \in B_{\ell,L}, D = d, Z = z\} \times \inf_{y\in B_{\ell,L}} f(y,d,z)$$
$$= \sum_{(\ell,d,z)\in\mathcal{M}_L} \sum_{(c_o,r_t,\chi)\in\mathcal{I}(\mathcal{R})} Q\{R_o \in \mathcal{Y}(c_o, r_t, \chi), R_t = r_t\} I\{c_o(d) = \ell, r_t(z) = d\} P\{Z = z\}$$
$$\times \inf_{y\in B_{\ell,L}} f(y,d,z) \ . \quad (112)$$

Results (111) and (112) together with $\text{diam}\{B_{\ell,L}\} < \delta$ and f satisfying (106) then imply $\int f(dP - d\tilde{P}) \leq \eta$.

Case II. Second, suppose that (108) holds with equality. Let $\bar{B}_{c_o}$ denote the closure of B_{c_o} and then note that $\bar{B}_{c_o} \cap \mathcal{R}'_o(r_t)$ is compact since $\mathcal{R}'_o(r_t)$ is closed by Assumption 5.1(iv) and $B_{\ell,L}$ is bounded. Since g is continuous, it follows that for each $(c_o, r_t) \in \mathcal{L}^{|\mathcal{D}|} \times \mathcal{R}_t$ such that $B_{c_o} \cap \mathcal{R}'_o(r_t) \neq \emptyset$, there is a $v^\star(c_o, r_t) \in \bar{B}_{c_o} \cap \mathcal{R}'(r_t)$ satisfying

$$\sup_{r_o \in B_{c_o}\cap\mathcal{R}'(r_t)} g(r_o, r_t) = g(v^\star(c_o, r_t), r_t) \ . \quad (113)$$

Note that $v^\star(c_o, r_t)$ does not necessarily belong to B_{c_0}. However, since $\{B_{\tilde{c}_o}\}_{\tilde{c}_o\in\mathcal{L}^{|\mathcal{D}|}}$ is a partition of $\mathcal{Y}^{|\mathcal{D}|}$ and $\mathcal{Y}^{|\mathcal{D}|}$ is closed, it follows that $v^\star(c_o, r_t) \in B_{\tilde{c}_o}$ for some $\tilde{c}_o$. Moreover, for any $d \in \mathcal{D}$ and $v^\star_d(c_o, r_t)$ denoting the d^{th} coordinate of $v^\star(c_o, r_t) \in \mathcal{Y}^{|\mathcal{D}|}$ we can conclude from the triangle inequality that

$$I\{v^\star(c_o, r_t) \in B_{\tilde{c}_o}\} \times \sup_{y\in B_{c_o(d),L}} \sup_{y'\in B_{\tilde{c}_o(d),L}} |y - y'|$$
$$\leq I\{v^\star(c_o, r_t) \in B_{\tilde{c}_o}\} \times \{ \sup_{y\in B_{c_o(d),L}} |y - v^\star_d(c_o, r_t)| + \sup_{y'\in B_{\tilde{c}_o(d),L}} |v^\star_d(c_o, r_t) - y'|\} \leq 2\delta \ , \quad (114)$$

where the final inequality follows from $v^\star_d(c_o, r_t)$ belonging to the closure of $B_{c_o(d),L}$ because $v^\star(c_o, r_t) \in \bar{B}_{c_o}$, $v^\star_d(c_o, r_t) \in B_{\tilde{c}_o(d),L}$, and $\text{diam}\{B_{l,L}\} < \delta$. Finally, we let $\tilde{Q}_R$ be a discrete measure with support points $(s(c_o, r_t, \chi) : (c_o, r_t, \chi) \in \mathcal{I}(\mathcal{R}))$ given by

$$s(c_o, r_t, \chi) = \begin{cases} (v^\star(c_o, r_t), r_t) & \text{if } \chi = 1 \\ \text{any } (r_o, r_t) \in (\mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}) \cap \mathcal{R} & \text{if } \chi = 0 \ , \end{cases}$$

and assign them probabilities $\tilde{Q}_R\{(R_o, R_t) = s(c_o, r_t, \chi)\} = Q_R\{(R_o, R_t) \in \mathcal{Y}(c_o, r_t, \chi) \times \{r_t\}\}$. Since $(v^\star(c_o, r_t), r_t) \in \mathcal{R}' \subseteq \mathcal{R}$ for any (c_o, r_t) such that $B_{c_o} \cap \mathcal{R}'_o(r_t) \neq \emptyset$, we then have that $\tilde{Q}_R\{R \in \mathcal{R}\} = 1$ due to Q_R satisfying (107), and in addition $E_{\tilde{Q}_R}[(g(R) - \theta_0)I\{R \in \mathcal{R}\}] = 0$ due to (113) and (108) holding with equality. Moreover, also note

$$\int f d\tilde{P} \geq \sum_{(\ell,d,z)\in\mathcal{M}_L} \sum_{(c_o,r_t)\in\mathcal{Y}^{|\mathcal{D}|}\times\mathcal{R}_t} \tilde{Q}\{R_o \in B_{c_o}, R_t = r_t\} I\{c_o(d) = \ell, r_t(z) = d\} P\{Z = z\} \times \inf_{y\in B_{\ell,L}} f(y, d, z) \quad (115)$$

by result (110) applied with $\tilde{P}$ and $\tilde{Q}$ in place of P and Q. In addition, by definition of $\tilde{Q}$ it follows that

$$\tilde{Q}\{R_o \in B_{c_o}, R_t = r_t\} = \sum_{\tilde{c}_o,\chi:(\tilde{c}_o,r_t,\chi)\in\mathcal{I}(\mathcal{R})} I\{s(\tilde{c}_o, r_t, \chi) \in B_{c_o} \times \{r_t\}\} Q\{R_o \in \mathcal{Y}(\tilde{c}_o, r_t, \chi), R_t = r_t\} ,$$

while result (114) together with the function $f : \mathcal{M} \to \mathbf{R}$ satisfying restriction (107) allow us to conclude that

$$\sum_{\ell=1}^{L} I\{c_o(d) = \ell, r_t(z) = d\} I\{s(\tilde{c}_o, r_t, \chi) \in B_{c_o} \times \{r_t\}\} \times \inf_{y\in B_{\ell,L}} f(y, d, z) \geq \sum_{\ell=1}^{L} I\{\tilde{c}_o(d) = \ell, r_t(z) = d\} I\{s(\tilde{c}_o, r_t, \chi) \in B_{\tilde{c}_o} \times \{r_t\}\} \times (\inf_{y\in B_{\ell,L}} f(y, d, z) - \eta) . \quad (116)$$

Hence, since $\sum_{c_o:(c_o,r_t,\chi)\in\mathcal{I}(\mathcal{R})} I\{s(\tilde{c}_o, r_t, \chi) \in B_{c_o} \times \{r_t\}\} = 1$, combining results (115)-(116) yields the bound

$$\int f d\tilde{P} \geq \sum_{(\ell,d,z)\in\mathcal{M}_L} \sum_{(\tilde{c}_o,r_t,\chi)\in\mathcal{I}(\mathcal{R})} Q\{R_o \in \mathcal{Y}(\tilde{c}_o, r_t, \chi), R_t = r_t\} I\{\tilde{c}_o(d) = \ell, r_t(z) = d\} P\{Z = z\} \times \left(\inf_{y\in B_{\ell,L}} f(y, d, z) - \eta\right) .$$

Therefore, the upper bound for $\int f dP$ obtained in (111), the function $f : \mathcal{M} \to \mathbf{R}$ satisfying (106), and $\text{diam}\{B_{\ell,L}\} < \delta$ for all $1 \leq \ell \leq L$ imply that $\int f(dP - \tilde{d}P) \leq 2\eta$.

Case III. The third case with (109) holding with equality follows from the same arguments from Case II. ■

Lemma E.12. *Suppose $\mathbf{P}_0 \neq \emptyset$, $\mathbf{P}$ is convex, and $\mathbf{Q}$ is the set of distributions for (R_o, R_t, Z) satisfying Assumptions 2.1-2.2. Also let $\text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ denote the closure of $\mathbf{P}_0$ under weak convergence (in $\mathbf{P}$) and $\tilde{P}_Z$ denote the marginal distribution of Z under $\tilde{P}$ for any $\tilde{P} \in \mathbf{P}$. If $P \in \mathbf{P} \setminus \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$, then there exist a continuous and bounded function f satisfying $\int f dP = 1$ and $\int f d\tilde{P} \leq 0$ for all $\tilde{P} \in \mathbf{P}_0$ satisfying $\tilde{P}_Z = P_Z$.*

Proof. Let $\mathbf{V}$ denote the linear span of $\mathbf{P}$, which forms a vector space, and define $C_b(\mathcal{M})$ to equal

$$C_b(\mathcal{M}) \equiv \{f : \mathcal{M} \to \mathbf{R} \text{ s.t. } f \text{ is continuous and bounded } \}$$

We equip $\mathbf{V}$ with the weakest topology that makes the linear maps $\tilde{P} \mapsto \int f \tilde{P}$ continuous for each $f \in C_b(\mathcal{M})$, which we denote by $\sigma(\mathbf{V}, C_b(\mathcal{M}))$. By Theorem 5.73 and Lemmas 2.52 and 2.53 in Aliprantis and Border (2006), $(\mathbf{V}, \sigma(\mathbf{V}, C_b(\mathcal{M})))$ is a locally convex topological vector space and the relative topology on $\mathbf{P}$ induced by $\sigma(\mathbf{V}, C_b(\mathcal{M}))$ equals the topology of weak convergence on $\mathbf{P}$. Next set $\mathbf{P}_0(P) \equiv \{\tilde{P} \in \mathbf{P}_0 : \tilde{P}_Z = P_Z\}$, let $\text{cl}_{\mathbf{P}}(\mathbf{P}_0(P))$ denote the closure of $\mathbf{P}_0(P)$ under weak convergence (in $\mathbf{P}$), and note that $\mathbf{P}_0(P) \subseteq \mathbf{P}_0$ implies $\text{cl}_{\mathbf{P}}(\mathbf{P}_0(P)) \subseteq \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ and hence $P \notin \text{cl}_{\mathbf{P}}(\mathbf{P}_0(P))$ because $P \notin \text{cl}_{\mathbf{P}}(\mathbf{P}_0)$ by hypothesis. Moreover, setting $\text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$ to denote the $\sigma(\mathbf{V}, C_b(\mathcal{M}))$-closure of $\mathbf{P}_0(P)$ in $\mathbf{V}$, it follows from Theorem 17.4 in Munkres (2000) that $\text{cl}_{\mathbf{P}}(\mathbf{P}_0(P)) = \text{cl}_{\mathbf{V}}(\mathbf{P}_0(P)) \cap \mathbf{P}$. Therefore, since $P \in \mathbf{P} \setminus \text{cl}_{\mathbf{P}}(\mathbf{P}_0(P))$, we can conclude that $P \notin \text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$. Also note that since $\mathbf{P}_0(P)$ is nonempty and convex by Lemma E.13, Lemma 5.27(6) in Aliprantis and Border (2006) implies $\text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$ is nonempty and convex as well.

We have so far shown that $P \notin \text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$ and that $\text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$ is a nonempty closed convex subset of the locally convex topological vector space $(\mathbf{V}, \sigma(\mathbf{V}, C_b(\mathcal{M}))$. By Corollary 5.80 and Theorem 5.93 in Aliprantis and Border (2006), there therefore exists a $\tilde{f} \in C_b(\mathcal{M})$ satisfying $\int \tilde{f} d\tilde{P} \leq 0$ for all $\tilde{P} \in \text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$ and $\int \tilde{f} dP > 0$. The lemma then follows by setting $f = \tilde{f} / \int \tilde{f} dP$ and using that $\mathbf{P}_0(P) \subseteq \text{cl}_{\mathbf{V}}(\mathbf{P}_0(P))$. ■

Lemma E.13. *Suppose $\mathbf{P}_0$ is nonempty, $\mathbf{P}$ is convex, $\mathbf{Q}$ is the set of all distributions for (R_o, R_t, Z) satisfying Assumptions 2.1-2.2, and define $\mathbf{P}_0(P) \equiv \{\tilde{P} \in \mathbf{P}_0 : \tilde{P}_Z = P_Z\}$ where $\tilde{P}_Z$ denotes the marginal of Z under $\tilde{P}_Z$. Then, $\mathbf{P}_0(P)$ is nonempty and convex.*

Proof. We first show $\mathbf{P}_0(P)$ is nonempty. Because $\mathbf{P}_0$ is nonempty, there exists Q that satisfies Assumptions 2.1–2.2 such that $Q\{R \in \mathcal{R}'\} > 0$ and $\theta(Q) = \theta_0$. By Assumption 2.1, $R \perp\!\!\!\perp Z$ under Q and therefore $Q = Q_R \times Q_Z$ where Q_R and Q_Z denote the marginal distributions of R and Z under Q. Next, define $\tilde{Q} = Q_R \times P_Z$ and note that $\tilde{Q} \in \mathbf{Q}$ because

$Q \in \mathbf{Q}$, while $\theta(\tilde{Q}) = \theta(Q) = \theta_0$ and $\tilde{Q}\{R \in \mathcal{R}'\} = Q\{R \in \mathcal{R}'\} > 0$ because $Q_R = \tilde{Q}_R$. Hence, setting $\tilde{P} = \tilde{Q}T^{-1}$ we obtain that $\tilde{P} \in \mathbf{P}_0(P)$, which shows that $\mathbf{P}_0(P)$ is nonempty.

Next, we show $\mathbf{P}_0(P)$ is convex. To this end, fix arbitrary $P^1, P^2 \in \mathbf{P}_0(P)$ and $\lambda \in [0,1]$. By definition of $\mathbf{P}_0(P)$, there then exist Q^1 and Q^2 that satisfy Assumptions 2.1–2.2 and $P^j = Q^j T^{-1}$, $Q^j\{R \in \mathcal{R}'\} > 0$, $\theta(Q^j) = \theta_0$, and $Q^j_Z = P_Z$ for $j \in \{1, 2\}$. Defining $Q = \lambda Q^1 + (1-\lambda)Q^2$, it then follows that

$$Q = \lambda Q^1_R \times P_Z + (1-\lambda)Q^2_R \times P_Z = (\lambda Q^1_R + (1-\lambda)Q^2_R) \times P_Z$$

and hence that Q satisfies Assumption 2.1 with $Q_Z = P_Z$. In addition, for any set $\mathcal{A}$ we have $Q\{R \in \mathcal{A}\} = \lambda Q^1\{R \in \mathcal{A}\} + (1-\lambda)Q^2\{R \in A\}$, which implies $Q\{R \in \mathcal{R}\} = 1$ and $Q\{R \in \mathcal{R}'\} > 0$. Moreover, we have

$$\theta(Q) = \frac{\lambda E_{Q^1}[g(R)I\{R \in \mathcal{R}'\}] + (1-\lambda)E_{Q^2}[g(R)I\{R \in \mathcal{R}'\}]}{\lambda Q^1\{R \in \mathcal{R}'\} + (1-\lambda)Q^2\{R \in \mathcal{R}'\}} = \frac{\lambda\theta_0 Q^1\{R \in \mathcal{R}'\} + (1-\lambda)\theta_0 Q^2\{R \in \mathcal{R}'\}}{\lambda Q^1\{R \in \mathcal{R}'\} + (1-\lambda)Q^2\{R \in \mathcal{R}'\}} = \theta_0$$

because $\theta(Q^1) = \theta(Q^2) = \theta_0$. Since $\lambda P^1 + (1-\lambda)P^2 = QT^{-1}$ and $\mathbf{P}$ is convex, it follows that $\lambda P^1 + (1-\lambda)P^2 \in \mathbf{P}_0(P)$, and therefore that $\mathbf{P}_0(P)$ is convex. ■

F Additional Simulation Details

F.1 Latent distribution under one-sided noncompliance

Table 6 presents the Q distribution used in the simulation in Section 6.1 and 6.2 that satisfies $Q \in \mathbf{Q}_{1s}$. For each row, the first four columns $Q(a_0, a_1, a_2, b_0, b_1, b_2)$ denote $Q\{Y(0) = a_0, Y(1) = a_1, Y(2) = a_2, D(0) = b_0, D(1) = b_1, D(2) = b_2\}$ in Section 6.1 or $\tilde{Q}\{\mu(0) = a_0, \mu(1) = a_1, \mu(2) = a_2, D(0) = b_0, D(1) = b_1, D(2) = b_2\}$ in Section 6.2, for $(a_0, a_1, a_2, b_0, b_1, b_2) \in \{-1, 0, 1\}^3 \times \{0, 1, 2\}^3$, and the last four columns represent the probabilities assigned to each of the four events respectively. The sum of the last four columns equals one.

F.2 Latent distribution under the encouragement design

Table 7 presents the Q distribution used in the simulation in Section 6.1 that satisfies $Q \in \mathbf{Q}_{enc}$. For each row, the first five columns $Q(a_0, a_1, a_2, b_0, b_1, b_2)$ denote $Q\{Y(0) = a_0, Y(1) =$

$a_1, Y(2) = a_2, D(0) = b_0, D(1) = b_1, D(2) = b_2\}$, for $(a_0, a_1, a_2, b_0, b_1, b_2) \in \{-1, 0, 1\}^3 \times \{0, 1, 2\}^3$, and the last five columns represent the probabilities assigned to each of the five events respectively. The sum of the last five columns equals one.

Events $(a_0, a_1, a_2, b_0, b_1, b_2)$				Probabilities assigned			
$Q(-1,-1,-1,0,0,0)$	$Q(-1,-1,-1,0,1,0)$	$Q(-1,-1,-1,0,0,2)$	$Q(-1,-1,-1,0,1,2)$	0.0001	0.0001	0.0001	0.0001
$Q(-1,-1,0,0,0,0)$	$Q(-1,-1,0,0,1,0)$	$Q(-1,-1,0,0,0,2)$	$Q(-1,-1,0,0,1,2)$	0.0001	0.0044	0.0067	0.0051
$Q(-1,-1,1,0,0,0)$	$Q(-1,-1,1,0,1,0)$	$Q(-1,-1,1,0,0,2)$	$Q(-1,-1,1,0,1,2)$	0.0001	0.0063	0.0001	0.0183
$Q(-1,0,-1,0,0,0)$	$Q(-1,0,-1,0,1,0)$	$Q(-1,0,-1,0,0,2)$	$Q(-1,0,-1,0,1,2)$	0.0001	0.0001	0.0009	0.0001
$Q(-1,0,0,0,0,0)$	$Q(-1,0,0,0,1,0)$	$Q(-1,0,0,0,0,2)$	$Q(-1,0,0,0,1,2)$	0.0001	0.0001	0.0070	0.0001
$Q(-1,0,1,0,0,0)$	$Q(-1,0,1,0,1,0)$	$Q(-1,0,1,0,0,2)$	$Q(-1,0,1,0,1,2)$	0.0001	0.0001	0.0105	0.0056
$Q(-1,1,-1,0,0,0)$	$Q(-1,1,-1,0,1,0)$	$Q(-1,1,-1,0,0,2)$	$Q(-1,1,-1,0,1,2)$	0.0001	0.0118	0.0001	0.0820
$Q(-1,1,0,0,0,0)$	$Q(-1,1,0,0,1,0)$	$Q(-1,1,0,0,0,2)$	$Q(-1,1,0,0,1,2)$	0.0001	0.0001	0.0001	0.0064
$Q(-1,1,1,0,0,0)$	$Q(-1,1,1,0,1,0)$	$Q(-1,1,1,0,0,2)$	$Q(-1,1,1,0,1,2)$	0.0001	0.0022	0.0001	0.0001
$Q(0,-1,-1,0,0,0)$	$Q(0,-1,-1,0,1,0)$	$Q(0,-1,-1,0,0,2)$	$Q(0,-1,-1,0,1,2)$	0.0045	0.0001	0.0012	0.0033
$Q(0,-1,0,0,0,0)$	$Q(0,-1,0,0,1,0)$	$Q(0,-1,0,0,0,2)$	$Q(0,-1,0,0,1,2)$	0.0077	0.0304	0.0021	0.0280
$Q(0,-1,1,0,0,0)$	$Q(0,-1,1,0,1,0)$	$Q(0,-1,1,0,0,2)$	$Q(0,-1,1,0,1,2)$	0.0001	0.0001	0.0004	0.0103
$Q(0,0,-1,0,0,0)$	$Q(0,0,-1,0,1,0)$	$Q(0,0,-1,0,0,2)$	$Q(0,0,-1,0,1,2)$	0.0001	0.0001	0.0001	0.0001
$Q(0,0,0,0,0,0)$	$Q(0,0,0,0,1,0)$	$Q(0,0,0,0,0,2)$	$Q(0,0,0,0,1,2)$	0.0001	0.0001	0.0001	0.0001
$Q(0,0,1,0,0,0)$	$Q(0,0,1,0,1,0)$	$Q(0,0,1,0,0,2)$	$Q(0,0,1,0,1,2)$	0.0001	0.0067	0.0001	0.4248
$Q(0,1,-1,0,0,0)$	$Q(0,1,-1,0,1,0)$	$Q(0,1,-1,0,0,2)$	$Q(0,1,-1,0,1,2)$	0.0001	0.0001	0.0001	0.0001
$Q(0,1,0,0,0,0)$	$Q(0,1,0,0,1,0)$	$Q(0,1,0,0,0,2)$	$Q(0,1,0,0,1,2)$	0.0159	0.0001	0.0001	0.0001
$Q(0,1,1,0,0,0)$	$Q(0,1,1,0,1,0)$	$Q(0,1,1,0,0,2)$	$Q(0,1,1,0,1,2)$	0.0001	0.0098	0.0001	0.0120
$Q(1,-1,-1,0,0,0)$	$Q(1,-1,-1,0,1,0)$	$Q(1,-1,-1,0,0,2)$	$Q(1,-1,-1,0,1,2)$	0.0001	0.0360	0.0001	0.0677
$Q(1,-1,0,0,0,0)$	$Q(1,-1,0,0,1,0)$	$Q(1,-1,0,0,0,2)$	$Q(1,-1,0,0,1,2)$	0.0008	0.0028	0.0003	0.0055
$Q(1,-1,1,0,0,0)$	$Q(1,-1,1,0,1,0)$	$Q(1,-1,1,0,0,2)$	$Q(1,-1,1,0,1,2)$	0.0033	0.0001	0.0033	0.0001
$Q(1,0,-1,0,0,0)$	$Q(1,0,-1,0,1,0)$	$Q(1,0,-1,0,0,2)$	$Q(1,0,-1,0,1,2)$	0.0056	0.0149	0.0001	0.0001
$Q(1,0,0,0,0,0)$	$Q(1,0,0,0,1,0)$	$Q(1,0,0,0,0,2)$	$Q(1,0,0,0,1,2)$	0.0001	0.0403	0.0083	0.0029
$Q(1,0,1,0,0,0)$	$Q(1,0,1,0,1,0)$	$Q(1,0,1,0,0,2)$	$Q(1,0,1,0,1,2)$	0.0008	0.0001	0.0001	0.0113
$Q(1,1,-1,0,0,0)$	$Q(1,1,-1,0,1,0)$	$Q(1,1,-1,0,0,2)$	$Q(1,1,-1,0,1,2)$	0.0001	0.0001	0.0001	0.0056
$Q(1,1,0,0,0,0)$	$Q(1,1,0,0,1,0)$	$Q(1,1,0,0,0,2)$	$Q(1,1,0,0,1,2)$	0.0001	0.0001	0.0227	0.0001
$Q(1,1,1,0,0,0)$	$Q(1,1,1,0,1,0)$	$Q(1,1,1,0,0,2)$	$Q(1,1,1,0,1,2)$	0.0049	0.0001	0.0001	0.0325

Table 6: Distribution of $Q \in \mathbf{Q}_{1s}$.

Events $(a_0, a_1, a_2, b_0, b_1, b_2)$					Probabilities assigned				
$Q(-1,-1,-1,0,0,0)$	$Q(-1,-1,-1,0,0,2)$	$Q(-1,-1,-1,0,1,0)$	$Q(-1,-1,-1,0,1,1)$	$Q(-1,-1,-1,0,1,2)$	0.0001	0.0001	0.0008	0.0002	0.0001
$Q(-1,-1,-1,0,2,2)$	$Q(-1,-1,-1,1,1,1)$	$Q(-1,-1,-1,1,1,2)$	$Q(-1,-1,-1,2,1,2)$	$Q(-1,-1,-1,2,2,2)$	0.0001	0.0023	0.0004	0.0001	0.0001
$Q(-1,-1,0,0,0,0)$	$Q(-1,-1,0,0,0,2)$	$Q(-1,-1,0,0,1,0)$	$Q(-1,-1,0,0,1,1)$	$Q(-1,-1,0,0,1,2)$	0.0001	0.0001	0.0001	0.0049	0.0012
$Q(-1,-1,0,0,2,2)$	$Q(-1,-1,0,1,1,1)$	$Q(-1,-1,0,1,1,2)$	$Q(-1,-1,0,2,1,2)$	$Q(-1,-1,0,2,2,2)$	0.0001	0.0001	0.0001	0.0037	0.0001
$Q(-1,-1,1,0,0,0)$	$Q(-1,-1,1,0,0,2)$	$Q(-1,-1,1,0,1,0)$	$Q(-1,-1,1,0,1,1)$	$Q(-1,-1,1,0,1,2)$	0.0001	0.0001	0.0024	0.0001	0.0631
$Q(-1,-1,1,0,2,2)$	$Q(-1,-1,1,1,1,1)$	$Q(-1,-1,1,1,1,2)$	$Q(-1,-1,1,2,1,2)$	$Q(-1,-1,1,2,2,2)$	0.0001	0.0001	0.0041	0.0001	0.0001
$Q(-1,0,-1,0,0,0)$	$Q(-1,0,-1,0,0,2)$	$Q(-1,0,-1,0,1,0)$	$Q(-1,0,-1,0,1,1)$	$Q(-1,0,-1,0,1,2)$	0.0003	0.0001	0.0001	0.0043	0.0001
$Q(-1,0,-1,0,2,2)$	$Q(-1,0,-1,1,1,1)$	$Q(-1,0,-1,1,1,2)$	$Q(-1,0,-1,2,1,2)$	$Q(-1,0,-1,2,2,2)$	0.0001	0.0001	0.0024	0.0013	0.0001
$Q(-1,0,0,0,0,0)$	$Q(-1,0,0,0,0,2)$	$Q(-1,0,0,0,1,0)$	$Q(-1,0,0,0,1,1)$	$Q(-1,0,0,0,1,2)$	0.0034	0.0166	0.0001	0.0059	0.0143
$Q(-1,0,0,0,2,2)$	$Q(-1,0,0,1,1,1)$	$Q(-1,0,0,1,1,2)$	$Q(-1,0,0,2,1,2)$	$Q(-1,0,0,2,2,2)$	0.0001	0.0001	0.0454	0.0001	0.0001
$Q(-1,0,1,0,0,0)$	$Q(-1,0,1,0,0,2)$	$Q(-1,0,1,0,1,0)$	$Q(-1,0,1,0,1,1)$	$Q(-1,0,1,0,1,2)$	0.0053	0.0025	0.0001	0.0001	0.0001
$Q(-1,0,1,0,2,2)$	$Q(-1,0,1,1,1,1)$	$Q(-1,0,1,1,1,2)$	$Q(-1,0,1,2,1,2)$	$Q(-1,0,1,2,2,2)$	0.0001	0.0060	0.0072	0.0001	0.0051
$Q(-1,1,-1,0,0,0)$	$Q(-1,1,-1,0,0,2)$	$Q(-1,1,-1,0,1,0)$	$Q(-1,1,-1,0,1,1)$	$Q(-1,1,-1,0,1,2)$	0.0022	0.0001	0.0001	0.0158	0.0001
$Q(-1,1,-1,0,2,2)$	$Q(-1,1,-1,1,1,1)$	$Q(-1,1,-1,1,1,2)$	$Q(-1,1,-1,2,1,2)$	$Q(-1,1,-1,2,2,2)$	0.0048	0.0001	0.0042	0.0001	0.0001
$Q(-1,1,0,0,0,0)$	$Q(-1,1,0,0,0,2)$	$Q(-1,1,0,0,1,0)$	$Q(-1,1,0,0,1,1)$	$Q(-1,1,0,0,1,2)$	0.0038	0.0001	0.0001	0.0001	0.0128
$Q(-1,1,0,0,2,2)$	$Q(-1,1,0,1,1,1)$	$Q(-1,1,0,1,1,2)$	$Q(-1,1,0,2,1,2)$	$Q(-1,1,0,2,2,2)$	0.0128	0.0001	0.0001	0.0043	0.0001
$Q(-1,1,1,0,0,0)$	$Q(-1,1,1,0,0,2)$	$Q(-1,1,1,0,1,0)$	$Q(-1,1,1,0,1,1)$	$Q(-1,1,1,0,1,2)$	0.0001	0.0001	0.0001	0.0126	0.0004
$Q(-1,1,1,0,2,2)$	$Q(-1,1,1,1,1,1)$	$Q(-1,1,1,1,1,2)$	$Q(-1,1,1,2,1,2)$	$Q(-1,1,1,2,2,2)$	0.0001	0.0001	0.0001	0.0048	0.0094
$Q(0,-1,-1,0,0,0)$	$Q(0,-1,-1,0,0,2)$	$Q(0,-1,-1,0,1,0)$	$Q(0,-1,-1,0,1,1)$	$Q(0,-1,-1,0,1,2)$	0.0001	0.0072	0.0001	0.0001	0.0079
$Q(0,-1,-1,0,2,2)$	$Q(0,-1,-1,1,1,1)$	$Q(0,-1,-1,1,1,2)$	$Q(0,-1,-1,2,1,2)$	$Q(0,-1,-1,2,2,2)$	0.0001	0.0003	0.0001	0.0639	0.0001
$Q(0,-1,0,0,0,0)$	$Q(0,-1,0,0,0,2)$	$Q(0,-1,0,0,1,0)$	$Q(0,-1,0,0,1,1)$	$Q(0,-1,0,0,1,2)$	0.0001	0.0483	0.0001	0.0012	0.0001
$Q(0,-1,0,0,2,2)$	$Q(0,-1,0,1,1,1)$	$Q(0,-1,0,1,1,2)$	$Q(0,-1,0,2,1,2)$	$Q(0,-1,0,2,2,2)$	0.0001	0.0001	0.0001	0.0001	0.0001
$Q(0,-1,1,0,0,0)$	$Q(0,-1,1,0,0,2)$	$Q(0,-1,1,0,1,0)$	$Q(0,-1,1,0,1,1)$	$Q(0,-1,1,0,1,2)$	0.0001	0.0001	0.0001	0.0001	0.0001
$Q(0,-1,1,0,2,2)$	$Q(0,-1,1,1,1,1)$	$Q(0,-1,1,1,1,2)$	$Q(0,-1,1,2,1,2)$	$Q(0,-1,1,2,2,2)$	0.0019	0.0001	0.0001	0.0001	0.0067
$Q(0,0,-1,0,0,0)$	$Q(0,0,-1,0,0,2)$	$Q(0,0,-1,0,1,0)$	$Q(0,0,-1,0,1,1)$	$Q(0,0,-1,0,1,2)$	0.0001	0.0275	0.0001	0.0030	0.0416
$Q(0,0,-1,0,2,2)$	$Q(0,0,-1,1,1,1)$	$Q(0,0,-1,1,1,2)$	$Q(0,0,-1,2,1,2)$	$Q(0,0,-1,2,2,2)$	0.0063	0.0001	0.0001	0.0001	0.0001
$Q(0,0,0,0,0,0)$	$Q(0,0,0,0,0,2)$	$Q(0,0,0,0,1,0)$	$Q(0,0,0,0,1,1)$	$Q(0,0,0,0,1,2)$	0.0078	0.0001	0.0038	0.0063	0.0001
$Q(0,0,0,0,2,2)$	$Q(0,0,0,1,1,1)$	$Q(0,0,0,1,1,2)$	$Q(0,0,0,2,1,2)$	$Q(0,0,0,2,2,2)$	0.0001	0.0047	0.0397	0.0001	0.0033
$Q(0,0,1,0,0,0)$	$Q(0,0,1,0,0,2)$	$Q(0,0,1,0,1,0)$	$Q(0,0,1,0,1,1)$	$Q(0,0,1,0,1,2)$	0.0001	0.0004	0.0001	0.0001	0.0053
$Q(0,0,1,0,2,2)$	$Q(0,0,1,1,1,1)$	$Q(0,0,1,1,1,2)$	$Q(0,0,1,2,1,2)$	$Q(0,0,1,2,2,2)$	0.0006	0.0001	0.0001	0.0019	0.0013
$Q(0,1,-1,0,0,0)$	$Q(0,1,-1,0,0,2)$	$Q(0,1,-1,0,1,0)$	$Q(0,1,-1,0,1,1)$	$Q(0,1,-1,0,1,2)$	0.0001	0.0017	0.0001	0.0028	0.0001
$Q(0,1,-1,0,2,2)$	$Q(0,1,-1,1,1,1)$	$Q(0,1,-1,1,1,2)$	$Q(0,1,-1,2,1,2)$	$Q(0,1,-1,2,2,2)$	0.0001	0.0001	0.0001	0.0001	0.0001
$Q(0,1,0,0,0,0)$	$Q(0,1,0,0,0,2)$	$Q(0,1,0,0,1,0)$	$Q(0,1,0,0,1,1)$	$Q(0,1,0,0,1,2)$	0.0008	0.0001	0.0001	0.0001	0.0010
$Q(0,1,0,0,2,2)$	$Q(0,1,0,1,1,1)$	$Q(0,1,0,1,1,2)$	$Q(0,1,0,2,1,2)$	$Q(0,1,0,2,2,2)$	0.0001	0.0011	0.0052	0.0023	0.0001
$Q(0,1,1,0,0,0)$	$Q(0,1,1,0,0,2)$	$Q(0,1,1,0,1,0)$	$Q(0,1,1,0,1,1)$	$Q(0,1,1,0,1,2)$	0.0053	0.0001	0.0035	0.0002	0.0008
$Q(0,1,1,0,2,2)$	$Q(0,1,1,1,1,1)$	$Q(0,1,1,1,1,2)$	$Q(0,1,1,2,1,2)$	$Q(0,1,1,2,2,2)$	0.0053	0.0001	0.0001	0.0022	0.0054
$Q(1,-1,-1,0,0,0)$	$Q(1,-1,-1,0,0,2)$	$Q(1,-1,-1,0,1,0)$	$Q(1,-1,-1,0,1,1)$	$Q(1,-1,-1,0,1,2)$	0.0001	0.0052	0.0001	0.0086	0.1091
$Q(1,-1,-1,0,2,2)$	$Q(1,-1,-1,1,1,1)$	$Q(1,-1,-1,1,1,2)$	$Q(1,-1,-1,2,1,2)$	$Q(1,-1,-1,2,2,2)$	0.0001	0.0014	0.0001	0.0011	0.0058
$Q(1,-1,0,0,0,0)$	$Q(1,-1,0,0,0,2)$	$Q(1,-1,0,0,1,0)$	$Q(1,-1,0,0,1,1)$	$Q(1,-1,0,0,1,2)$	0.0001	0.0001	0.0068	0.0015	0.0001
$Q(1,-1,0,0,2,2)$	$Q(1,-1,0,1,1,1)$	$Q(1,-1,0,1,1,2)$	$Q(1,-1,0,2,1,2)$	$Q(1,-1,0,2,2,2)$	0.0001	0.0001	0.0093	0.0029	0.0001
$Q(1,-1,1,0,0,0)$	$Q(1,-1,1,0,0,2)$	$Q(1,-1,1,0,1,0)$	$Q(1,-1,1,0,1,1)$	$Q(1,-1,1,0,1,2)$	0.0001	0.0022	0.0513	0.0175	0.0001
$Q(1,-1,1,0,2,2)$	$Q(1,-1,1,1,1,1)$	$Q(1,-1,1,1,1,2)$	$Q(1,-1,1,2,1,2)$	$Q(1,-1,1,2,2,2)$	0.0001	0.0023	0.0001	0.0001	0.0008
$Q(1,0,-1,0,0,0)$	$Q(1,0,-1,0,0,2)$	$Q(1,0,-1,0,1,0)$	$Q(1,0,-1,0,1,1)$	$Q(1,0,-1,0,1,2)$	0.0001	0.0001	0.0001	0.0041	0.0003
$Q(1,0,-1,0,2,2)$	$Q(1,0,-1,1,1,1)$	$Q(1,0,-1,1,1,2)$	$Q(1,0,-1,2,1,2)$	$Q(1,0,-1,2,2,2)$	0.0006	0.0001	0.0021	0.0155	0.0028
$Q(1,0,0,0,0,0)$	$Q(1,0,0,0,0,2)$	$Q(1,0,0,0,1,0)$	$Q(1,0,0,0,1,1)$	$Q(1,0,0,0,1,2)$	0.0001	0.0162	0.0020	0.0001	0.0001
$Q(1,0,0,0,2,2)$	$Q(1,0,0,1,1,1)$	$Q(1,0,0,1,1,2)$	$Q(1,0,0,2,1,2)$	$Q(1,0,0,2,2,2)$	0.0081	0.0001	0.0033	0.0001	0.0001
$Q(1,0,1,0,0,0)$	$Q(1,0,1,0,0,2)$	$Q(1,0,1,0,1,0)$	$Q(1,0,1,0,1,1)$	$Q(1,0,1,0,1,2)$	0.0090	0.0001	0.0001	0.0001	0.0001
$Q(1,0,1,0,2,2)$	$Q(1,0,1,1,1,1)$	$Q(1,0,1,1,1,2)$	$Q(1,0,1,2,1,2)$	$Q(1,0,1,2,2,2)$	0.0001	0.0068	0.0056	0.0001	0.0043
$Q(1,1,-1,0,0,0)$	$Q(1,1,-1,0,0,2)$	$Q(1,1,-1,0,1,0)$	$Q(1,1,-1,0,1,1)$	$Q(1,1,-1,0,1,2)$	0.0008	0.0001	0.0019	0.0025	0.0001
$Q(1,1,-1,0,2,2)$	$Q(1,1,-1,1,1,1)$	$Q(1,1,-1,1,1,2)$	$Q(1,1,-1,2,1,2)$	$Q(1,1,-1,2,2,2)$	0.0017	0.0001	0.0001	0.0001	0.0001
$Q(1,1,0,0,0,0)$	$Q(1,1,0,0,0,2)$	$Q(1,1,0,0,1,0)$	$Q(1,1,0,0,1,1)$	$Q(1,1,0,0,1,2)$	0.0001	0.0050	0.0001	0.0001	0.0040
$Q(1,1,0,0,2,2)$	$Q(1,1,0,1,1,1)$	$Q(1,1,0,1,1,2)$	$Q(1,1,0,2,1,2)$	$Q(1,1,0,2,2,2)$	0.0097	0.0001	0.0038	0.0017	0.0001
$Q(1,1,1,0,0,0)$	$Q(1,1,1,0,0,2)$	$Q(1,1,1,0,1,0)$	$Q(1,1,1,0,1,1)$	$Q(1,1,1,0,1,2)$	0.0001	0.0001	0.0050	0.0001	0.0001
$Q(1,1,1,0,2,2)$	$Q(1,1,1,1,1,1)$	$Q(1,1,1,1,1,2)$	$Q(1,1,1,2,1,2)$	$Q(1,1,1,2,2,2)$	0.0062	0.0001	0.0016	0.0040	0.0107

Table 7: Distribution of $Q \in \mathbf{Q}_{\text{enc}}$.