Inference for Interval-Identified Parameters Selected from an Estimated Set[*]

Sukjin Han[†]   Adam McCloskey[‡]

February 29, 2024

Abstract

Interval identification of parameters such as average treatment effects, average partial effects and welfare is particularly common when using observational data and experimental data with imperfect compliance due to the endogeneity of individuals' treatment uptake. In this setting, a treatment or policy will typically become an object of interest to the researcher when it is either selected from the estimated set of best-performers or arises from a data-dependent selection rule. In this paper, we develop new inference tools for interval-identified parameters chosen via these forms of selection. We develop three types of confidence intervals for data-dependent and interval-identified parameters, discuss how they apply to several examples of interest and prove their uniform asymptotic validity under weak assumptions.

Keywords: Partial Identification, Post-Selection Inference, Conditional Inference, Uniform Validity, Treatment Choice.

1 Introduction

There is now a large and growing literature on partial identification of optimal treatments and policies under practically-relevant assumptions for observational data and experimental data with imper-

fect compliance (e.g., Stoye, 2012; Kallus and Zhou, 2021; Pu and Zhang, 2021; D'Adamo, 2021; Yata, 2021; Han, 2024 to name just a few). Interval identification of average treatment effects (ATEs), average partial effects and welfare is particularly common in these settings due to the endogeneity of individuals' treatment uptake. In order to use the partial identification results for treatment or policy choice in practice, a researcher must typically estimate a set of best-performing treatments or policies from data. Consequently, a treatment or policy will typically become an object of interest to the researcher when it is either selected from the estimated set of best-performers or arises from a data-dependent selection rule. It is now well-known that selecting an object of interest from data invalidates standard inference tools (e.g., Andrews et al., 2024). The failure of standard inference tools after data-dependent selection is only compounded by the presence of partially-identified parameters.

In this paper, we develop new inference tools for interval-identified parameters corresponding to selection from either an estimated set or arising from a data-dependent selection rule. Estimating identified sets for the best-performing treatments/policies or forming data-dependent selection rules in these settings is important for choosing which treatments/policies to implement in practice. Therefore, the ability to infer how well these treatments or policies should be expected to perform when selected, for instance to gauge whether their implementation is worthwhile, is of primary practical importance.

The current literature has not yet developed valid post-selection inference tools in partially-identified contexts, an important deficiency that the methods proposed in this paper aim to correct. The methods we propose here build upon the ideas of conditional and hybrid inference employed in various point-identified contexts by, e.g., Lee et al. (2016), Andrews et al. (2024) and McCloskey (2024) to produce confidence intervals (CIs) for interval-identified parameters such as welfare or ATEs chosen from an estimated set or via a data-dependent selection rule. Although Andrews et al. (2023) also propose conditional and hybrid inference methods in the partial identification context of moment inequality models, they do not allow for data-dependent selection of objects of interest, one of the main focuses of the present paper. Finally, this paper directly relies upon results in the literature on interval identification of welfare, treatment effects and partial outcomes such as Manski (1990), Balke and Pearl (1997, 2011), Manski and Pepper (2000), Han and Yang (2024) and

Han (2024). We apply our inference methods to a general class of problems nesting these examples.

After sketching the ideas behind our inference methods in a simple example, we introduce the general inference framework to which our methods can be applied. We show that our general inference framework incorporates several problems of interest for data-dependent selection and treatment rules for parameters belonging to an identified set such as Manski bounds for average potential outcomes or ATEs, bounds on parameters derived from linear programming (e.g., Balke and Pearl, 1997, 2011; Mogstad et al., 2018; Han, 2024; Han and Yang, 2024), bounds on welfare for treatment allocation rules that are partially identified by observational data (e.g., Stoye, 2012; Kallus and Zhou, 2021; Pu and Zhang, 2021; D'Adamo, 2021; Yata, 2021) and bounds for dynamic treatment effects (e.g., Han, 2024). We also show how to incorporate inference on parameters chosen via asymptotically optimal treatment choice rules (e.g., Christensen et al., 2023) into our general inference framework. Our framework can also be applied to settings where welfare is partially identified for other reasons (e.g., Cui and Han, 2024).

Within the general inference framework, we develop three types of CIs for data-dependent and interval-identified parameters. As the name suggests, the *conditional* CIs are asymptotically valid conditional on the parameter of interest corresponding to a treatment or policy chosen from an estimated set. The construction of this CI does not require a specific rule for choosing the parameter of interest from the estimated set. In addition, the sampling framework underlying its conditional validity is most appropriate in contexts for which a researcher will only be interested in the parameter because it is chosen from the estimated set. Importantly, we show that these CIs are asymptotically valid *uniformly* across a large class of data-generating processes (DGPs). Uniform asymptotic validity is especially important for approximately correct finite-sample coverage in post-selection contexts like those in this paper (see, e.g., Andrews and Guggenberger, 2009).

The second and third types of CIs we develop in this paper are designed for inference on parameters chosen by data-dependent selection rules for which the object of interest is uniquely determined by the selection rule. The *projection* CIs do not require knowledge of the selection rule to be asymptotically valid whereas the *hybrid* CI construction utilizes the particular form of a selection rule to improve upon the length properties of the projection CI. The conditional CIs are

short for selections that occur with high probability but can become exceptionally long when this probability is small (see, e.g., Kivaranovic and Leeb, 2021). Conversely, projection CIs are overly conservative when selection probabilities are high. Hybrid CIs interpolate the length properties of the conditional and projection CIs in order to attain good length properties regardless of the value selection probabilities take. In analogy with the conditional CIs, we formally show that both projection and hybrid CIs have are asymptotically valid in a uniform sense.

We analyze the coverage and length properties of our proposed CIs in finite samples. Since, to our knowledge, these are the first uniformly valid CIs for data-dependent selections of partially-identified parameters, there are no existing CIs to which we can direclty compare. Nevertheless, since our CIs can also be used for inference on a priori chosen interval-identified parameters, we conduct a power comparison with one of the leading methods for inference on a partially-identified parameter. In particular, we compare the power of a test implied by our hybrid CIs to the power of the hybrid test of Andrews et al. (2023), a test that applies to a general class of moment-inequality models that is also based upon a (different) hybrid between conditional and projection-based inference. Encouragingly, the power of the test implied by our hybrid CI is quite competitive even in this environment for which it was not designed. We also find that the finite-sample coverage of all of our CIs is approximately correct in a simple Manski bound example. Finally, we analyze the length tradeoffs between the three different CIs across different DGPs, finding the hybrid CI to perform best overall.

The remainder of this paper is structured as follows. Section 2 sketches the ideas behind our general CI constructions in the context of a simple Manski bound example. Section 3 lays down the general high-level inference framework we are interested in while Section 4 details how the general framework applies in several different examples. Sections 5 and 6 then detail the various CI constructions in the general setting. The final two sections of the paper, Sections 7 and 8, are devoted to finite-sample comparisons of the properties of the different CIs in the context of a simple Manski bound example. Mathematical proofs are relegated to a Technical Appendix at the end of the paper.

2 Basic Ideas: Inference with Manski Bounds Example

Consider a binary outcome of interest Y, a binary treatment indicator D and a binary treatment assignment Z. Furthermore, let $Y(0)$ and $Y(1)$ denote potential outcomes under no treatment ($D = 0$) and treatment ($D=1$). Assuming $E[Y(d)|Z] = E[Y(d)]$, Manski (1990) shows that we can bound the average potential outcomes $W(d) = \mathbb{E}[Y(d)]$ in the absence and presence of treatment as follows:

$$L(d) \equiv \max\{p^{1d0}, p^{1d1}\} \leq W(d) \leq \min\{1 - p^{0d0}, 1 - p^{0d1}\} \equiv U(d) \qquad (2.1)$$

for $d = 0,1$ and $p^{ydz} = Pr(Y = y, D = d|Z = z)$. Given (2.1) it is natural to estimate the set of best-performing options $\mathcal{D}^*$, as a subset of the two options of treatment and no treatment, from an observed dataset of outcomes, treatments and treatment assignments $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ as those that are estimated to be undominated:

$$\widehat{\mathcal{D}} = \left\{ d \in \{0,1\} : \widehat{U}(d) \geq \widehat{L}(d') \ \forall d' \in \{0,1\} \text{ s.th. } d' \neq d \right\},$$

where $\widehat{L}(d) \equiv \max\{\hat{p}^{1d0}, \hat{p}^{1d1}\}$ and $\widehat{U}(d) \equiv \min\{1 - \hat{p}^{0d0}, 1 - \hat{p}^{0d1}\}$ with $\hat{p}^{ydz}$ being an empirical estimate of $p^{ydz} = Pr(Y = y, D = d|Z = z)$.

We are interested in inference on the identified interval $[L(d), U(d)]$ for the average potential outcome $W(d)$ of option $d \in \{0,1\}$ after the researcher selects this option from $\widehat{\mathcal{D}}$. In other words, we would like to provide statistically precise statements about the true average potential outcome of an option selected from the data to give the researcher an idea of how well this selected option should be expected to perform in the population. We first provide some intuition for why standard inference techniques based upon asymptotic normality fail and then sketch our proposals for valid inference in this context.

2.1 Why Does Standard Inference Fail?

To fix ideas, let us focus for now on inference for the lower bound $L(d)$ of a selected option rather than the entire identified set $[L(d), U(d)]$. More specifically, since $L(d)$ is a lower bound for the av-

erage potential outcome $W(d)$, we would like to obtain a probabilistic lower bound for $L(d)$. Under standard conditions, a central limit theorem implies $\hat{p} = (\hat{p}^{100}, \hat{p}^{010}, \hat{p}^{110}, \hat{p}^{101}, \hat{p}^{011}, \hat{p}^{111})'$ is normally distributed in large samples. So why not form a CI using $\widehat{L}(d)$ and quantiles from a normal distribution as the basis for inference? There are two reasons such an approach is (asymptotically) invalid:

1. Even in the absence of selection, $\widehat{L}(d) = \max\{\hat{p}^{1d0}, \hat{p}^{1d1}\}$ is not normally distributed in large samples.

2. Data-dependent selection of d further complicates the distribution of $\widehat{L}(d)$.

Reason 1. is easy to see since $\widehat{L}(d)$ is the maximum of two normally distributed random variables in large samples when d is chosen a priori. To better understand reason 2., note that the distribution of $\widehat{L}(d)$ given $d \in \widehat{\mathcal{D}}$ is the conditional distribution of the maximum of two normally distributed random variables given that the minimum of two other normally distributed random variables, $\widehat{U}(d) \equiv \min\{1 - \hat{p}^{0d0}, 1 - \hat{p}^{0d1}\}$, exceeds the maximum of yet another set of two normally distributed random variables, $\widehat{L}(d') = \max\{\hat{p}^{1d'0}, \hat{p}^{1d'1}\}$ for $d' \neq d$. Unconditionally, $\widehat{L}(\hat{d})$ for any data-dependent choice of $\hat{d} \in \widehat{\mathcal{D}}$ is distributed as a mixture of the distributions of $\widehat{L}(0)$ and $\widehat{L}(1)$, neither of which are themselves normally distributed.

2.2 Conditional Confidence Intervals

Suppose that a researcher's interest in inference on $L(d)$ only arises when d is estimated to be in the set of best-performing options, viz., $d \in \widehat{\mathcal{D}}$. In such a case, we are interested in a probabilistic lower bound for $L(d)$ that is approximately valid across repeated samples for which $d \in \widehat{\mathcal{D}}$, i.e., we would like to form a *conditionally* valid lower bound $\widehat{L}(d)_\alpha^C$ such that[1]

$$\mathbb{P}\left(L(d) \geq \widehat{L}(d)_\alpha^C \,\Big|\, d \in \widehat{\mathcal{D}} \right) \geq 1 - \alpha \tag{2.2}$$

---

[1]See Andrews et al. (2024) for an extensive discussion of when conditional vs unconditional validity is desirable for inference after selection.

for some $\alpha \in (0,1)$ in large samples. To do so we characterize the conditional distribution of $\widehat{L}(d)$. Specifically, let $\hat{j}_L(d) = \text{argmax}_{j \in \{0,1\}} \hat{p}^{1dj}$. Lemma 5.1 of Lee et al. (2016) implies

$$\widehat{L}(d) \Big| \Big\{ d \in \widehat{\mathcal{D}}, \hat{j}_L(d) = j_L^* \Big\} \sim \xi_L \Big| \Big\{ \widehat{\mathcal{L}}_L(Z_L) \leq \xi_L \leq \widehat{\mathcal{U}}_L(Z_L) \Big\}, \tag{2.3}$$

for some known functions $\widehat{\mathcal{L}}_L(\cdot)$ and $\widehat{\mathcal{U}}_L(\cdot)$, where $\xi_L \sim \mathcal{N}(p^{1dj_L^*}, \text{Var}(\hat{p}^{1dj_L^*}))$ in large samples and Z_L is a sufficient statistic for the nuisance parameter $p = (p^{100}, p^{010}, p^{110}, p^{101}, p^{011}, p^{111})'$ that is asymptotically independent of ξ_L. Using results in Pfanzagl (1994), the characterization in (2.3) permits the straightforward computation of a conditionally quantile-unbiased estimator for $p^{1d\hat{j}_L(d)}$, $\widehat{L}(d)_\alpha^C$, satisfying

$$\mathbb{P}\Big(p^{1d\hat{j}_L(d)} \geq \widehat{L}(d)_\alpha^C \Big| d \in \widehat{\mathcal{D}} \Big) = 1 - \alpha \tag{2.4}$$

in large samples. However, noting that $L(d) \geq p^{1d\hat{j}_L(d)}$ with probability one, we can see that (2.2) holds for this choice of $\widehat{L}(d)_\alpha^C$.

Although (2.2) does not hold with exact equality, we note that the left-hand side cannot be much larger than the right-hand side. In other words, although $\widehat{L}(d)_\alpha^C$ is a conservative probabilistic lower bound for $L(d)$, it is not very conservative. This can be seen heuristically by working through the two possible values that $L(d)$ can take:

1. If $L(d) = p^{1d\hat{j}_L(d)}$, then (2.2) holds with equality in large samples by (2.4).

2. If $L(d) \neq p^{1d\hat{j}_L(d)}$, then $L(d) \approx p^{1d\hat{j}_L(d)}$ since $\widehat{L}(d) = \hat{p}^{1d\hat{j}_L(d)}$ so that the left-hand side of (2.2) cannot be much larger than the right-hand side.

Finally, an analogous construction to that described above for producing a probabilistic lower bound for $L(d)$ produces a conditionally valid probabilistic upper bound $\widehat{U}(d)_{1-\alpha}^C$ for $U(d)$ that satisfies

$$\mathbb{P}\Big(U(d) \leq \widehat{U}(d)_{1-\alpha}^C \Big| d \in \widehat{\mathcal{D}} \Big) \geq 1 - \alpha \tag{2.5}$$

for some $\alpha \in (0,1)$ in large samples. The probabilistic lower and upper bounds can then be combined

to form a CI, $[\widehat{L}(d)^C_{\alpha/2},\widehat{U}(d)^C_{1-\alpha/2}]$, that is conditionally valid for $[L(d),U(d)]$ in large samples since

$$\mathbb{P}\Big(L(d)\geq\widehat{L}(d)^C_{\alpha/2},U(d)\leq\widehat{U}(d)^C_{1-\alpha/2}\Big|d\in\widehat{\mathcal{D}}\Big) \tag{2.6}$$

$$\geq 1-\mathbb{P}\Big(L(d)<\widehat{L}(d)^C_{\alpha/2}\Big|d\in\widehat{\mathcal{D}}\Big)-\mathbb{P}\Big(U(d)>\widehat{U}(d)^C_{1-\alpha/2}\Big|d\in\widehat{\mathcal{D}}\Big)\geq 1-\alpha.$$

2.3 Unconditional Confidence Intervals

Suppose now that the researcher uses a data-dependent rule to select a unique option of inferential interest. For example, suppose the researcher is interested in choosing the option with the highest potential outcome in the worst case across its identified set so that she chooses $\hat{d}=\mathrm{argmax}_{d\in\{0,1\}}\widehat{L}(d)$. In such a case, it is natural to form a probabilistic lower bound $\widehat{L}(\hat{d})^U_\alpha$ for $L(\hat{d})$ that is *unconditionally* valid across repeated samples such that

$$\mathbb{P}\Big(L(\hat{d})\geq\widehat{L}(\hat{d})^U_\alpha\Big)\geq 1-\alpha \tag{2.7}$$

for some $\alpha\in(0,1)$ in large samples. Given its conditional validity (2.2), the conditional lower bound $\widehat{L}(d)^C_\alpha$ also satisfies (2.7) upon changing the definition of $\widehat{\mathcal{D}}$ to $\widehat{\mathcal{D}}=\{\hat{d}\}$ in its construction. However, it is well known in the literature on selective inference that conditionally-valid probabilistic bounds can be very uninformative (i.e., far below the true value) when the probability of the conditioning event is small (see e.g., Kivaranovic and Leeb, 2021, Andrews et al., 2024 and McCloskey, 2024). Here, we propose two additional forms of probabilistic bounds that are only unconditionally valid but do not suffer from this drawback.

First, we can form a probabilistic lower bound for $L(\hat{d})$ by projecting a one-sided rectangular simultaneous confidence lower bound for all possible values $L(\hat{d})$ can take: $\widehat{L}(\hat{d})^P_\alpha = \widehat{L}(\hat{d})-\sqrt{\widehat{\Sigma}_{L,2d+1+\hat{j}_L(d)}}\hat{c}_{1-\alpha,L}$, where $\hat{c}_{1-\alpha,L}$ is the $1-\alpha$ quantile of $\max_i\hat{\zeta}_i/\sqrt{\widehat{\Sigma}_{L,i}}$ for $\hat{\zeta}\sim\mathcal{N}(0,\widehat{\Sigma}_L)$, $\widehat{\Sigma}_L$ is a consistent estimator of $\Sigma_L=\mathrm{Var}(\hat{p}^{100},\hat{p}^{101},\hat{p}^{110},\hat{p}^{111})$ and $\mathcal{M}_i$ denotes the ith element of the main diagonal of the matrix $\mathcal{M}$. Since

$$\mathbb{P}\bigg((p^{100},p^{101},p^{110},p^{111})\geq(\hat{p}^{100},\hat{p}^{101},\hat{p}^{110},\hat{p}^{111})-\sqrt{\mathrm{Diag}(\widehat{\Sigma}_L)}\hat{c}_{1-\alpha,L}\bigg)=1-\alpha,$$

in large samples, $\mathbb{P}(p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\alpha^P) \geq 1-\alpha$ and (2.7) holds for $\widehat{L}(\hat{d})_\alpha^U = \widehat{L}(\hat{d})_\alpha^P$ becasue $L(\hat{d}) \geq p^{1\hat{d}\hat{j}_L(\hat{d})}$. However, $\widehat{L}(\hat{d})_\alpha^P$ suffers from a converse drawback to that of $\widehat{L}(\hat{d})_\alpha^C$: it is unnecessarily conservative when $\hat{d}=d$ is chosen with high probability (see e.g., Andrews et al., 2024 and McCloskey, 2024).

We propose a second probabilisitc lower bound for $L(\hat{d})$ that combines the complementary strengths of $\widehat{L}(d)_\alpha^C$ and $\widehat{L}(\hat{d})_\alpha^P$. Construction of this hybrid lower bound $\widehat{L}(\hat{d})_\alpha^H$ proceeds analogously to the construction of $\widehat{L}(d)_\alpha^C$ after adding the additional condition $\{p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\}$ for $\beta < \alpha$ to the conditioning event and instead computing a conditionally quantile-unbiased estimator for $p^{1\hat{d}\hat{j}_L(\hat{d})}$, $\widehat{L}(\hat{d})_\alpha^H$, satisfying

$$\mathbb{P}\left(p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\alpha^H \,\Big|\, \hat{d}=d^*, p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\right) = \frac{1-\alpha}{1-\beta}$$

in large samples, where d^* is any realized value of the random variable $\hat{d}$. Imposing this additional condition in the formation of the hybrid bound ensures that $\widehat{L}(\hat{d})_\alpha^H$ is always greater than $\widehat{L}(\hat{d})_\beta^P$, limiting its worst-case performance relative to $L(\hat{d})_\beta^P$ when $\mathbb{P}(\hat{d}=d^*)$ is small. On the other hand, when $\mathbb{P}(\hat{d}=d^*)$ is large, the additional condition $\{p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\}$ is far from binding with high probability so that $\widehat{L}(\hat{d})_\alpha^H$ becomes very close to $\widehat{L}(d)_{(\alpha-\beta)/(1-\beta)}^C$. In this case, $\widehat{L}(d)_{(\alpha-\beta)/(1-\beta)}^C$ is close to the naive lower bound based upon the normal distribution $\widehat{L}(\hat{d}) - \sqrt{\mathrm{Var}(\hat{p}^{1\hat{d}\hat{j}_L^*})}z_{(1-\alpha)/(1-\beta)}$ because the truncation bounds in (2.3) are very wide (Proposition 3 in Andrews et al., 2024).

To see how (2.7) holds for $\widehat{L}(\hat{d})_\alpha^U = \widehat{L}(\hat{d})_\alpha^H$, note first that

$$\mathbb{P}\left(L(\hat{d}) \geq \widehat{L}(\hat{d})_\alpha^H \,\Big|\, \hat{d}=d^*, p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\right)$$
$$\geq \mathbb{P}\left(p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\alpha^H \,\Big|\, \hat{d}=d^*, p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\right) = \frac{1-\alpha}{1-\beta}$$

for all $d^* \in \{0,1\}$. Then, note that

$$\mathbb{P}\left(L(\hat{d}) \geq \widehat{L}(\hat{d})_\alpha^H\right) \geq \mathbb{P}\left(L(\hat{d}) \geq \widehat{L}(\hat{d})_\alpha^H \,\Big|\, p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\right) \cdot \mathbb{P}\left(p^{1\hat{d}\hat{j}_L(\hat{d})} \geq \widehat{L}(\hat{d})_\beta^P\right)$$
$$\geq \frac{1-\alpha}{1-\beta}(1-\beta) = 1-\alpha$$

by the law of total probability.

By similar reasoning to that used for the conditional confidence intervals in Section 2.2 above, $\widehat{L}(\hat{d})^H_\alpha$ is not very conservative as a probabilistic lower bound for $L(\hat{d})$. The researcher's choice of $\beta \in (0,\alpha)$ trades off the performance of $\widehat{L}(\hat{d})^H_\alpha$ across scenarios for which $\mathbb{P}(\hat{d}=d^*)$ is large and small with a small β corresponding to better performance when $\mathbb{P}(\hat{d}=d^*)$ is large. See McCloskey (2024) for an in-depth discussion of these tradeoffs. We recommend $\beta = \alpha/10$.

Finally, analogous constructions to those above produce unconditional projection and hybrid probabilistic upper bounds $\widehat{U}(\hat{d})^P_{1-\alpha}$ and $\widehat{U}(\hat{d})^H_{1-\alpha}$ that can then be combined with the lower bounds to form CIs $[\widehat{L}(d)^P_{\alpha/2}, \widehat{U}(d)^P_{1-\alpha/2}]$ and $[\widehat{L}(d)^H_{\alpha/2}, \widehat{U}(d)^H_{1-\alpha/2}]$ for $[L(d),U(d)]$ that are unconditionally valid in large samples by the same arguments as those used in (2.6) above.

3   General Inference Framework

We now introduce the general inference framework that we are interested in, nesting the Manski bound example of the previous section as a special case. After introducing the general framework, we describe several additional example applications that fall within this framework.

We are interested in performing inference on a parameter $W(d)$ that is indexed by a finite set $d \in \mathcal{D} \equiv \{0,...,T\}$ for some $T>0$. The index d may correspond to a particular treatment, treatment allocation rule or policy, depending upon the application. We assume that $W(d)$ belongs to an identified set taking a particular interval form that is common to many applications of interest.

Assumption 3.1. *For all $d \in \{0,...,T\}$ and an unknown finite-dimensional parameter p,*

1. *$L(d) \equiv \max_{j \in \{1,...,J_L\}} \{\tilde{\ell}_{d,j} + \ell_{d,j} p\} \leq W(d)$ for some fixed and known J_L, $\tilde{\ell}_{d,1},...,\tilde{\ell}_{d,J_L}$ and nonzero $\ell_{d,1},...,\ell_{d,J_L}$ such that $\ell_{d,j} \neq \ell_{d,j'}$ for $j \neq j'$.*

2. *$U(d) \equiv \min_{j \in \{1,...,J_U\}} \{\tilde{u}_{d,j} + u_{d,j} p\} \geq W(d)$ for some fixed and known J_U, $\tilde{u}_{d,1},...,\tilde{u}_{d,J_U}$ and nonzero $u_{d,1},...,u_{d,J_U}$ such that $u_{d,j} \neq u_{d,j'}$ for $j \neq j'$.*

The lower and upper endpoints of identified sets for the welfare, average potential outcome or ATE of a treatment d typically take the form of $L(d)$ and $U(d)$ in cases for which (sequences of) outcomes, treatments and instruments are discrete.

In the setting of this paper, a researcher's interest in $W(d)$ arises when d belongs to a set $\widehat{\mathcal{D}} \subset \{0,...,T\}$ that is estimated from a sample of n observations. It is often the case that $\widehat{\mathcal{D}}$ is an estimate of $\mathcal{D}^*$. This set could correspond to an estimated set of optimal treatments or policies or other data-dependent index sets of interest. The estimated set is determined by an estimator $\hat{p}$ of the finite-dimensional parameter p that determines the bounds on $W(d)$ according to Assumption 3.1. Let

$$\hat{j}_L(d) = \underset{j \in \{1,...,J_L\}}{\operatorname{argmax}} \{\tilde{\ell}_{d,j} + \ell_{d,j}\hat{p}\}, \quad \hat{j}_U(d) = \operatorname{argmin}_{j \in \{1,...,J_U\}} \{\tilde{u}_{d,j} + u_{d,j}\hat{p}\}$$

so that the estimated lower and upper bounds for $W(d)$ are equal to $\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}\hat{p}$ and $\tilde{u}_{d,\hat{j}_U(d)} + u_{d,\hat{j}_U(d)}\hat{p}$. We work under the high-level assumption that the event that (i) a parameter is in the set of interest, (ii) the estimated bounds on $W(d)$ are realized at a given value and (iii) an additional random vector is realized at any given value can be written as a polyhedron in $\hat{p}$.

Assumption 3.2. *1. For some finite-valued random vector $\hat{\gamma}_L(d)$, some fixed and known matrix $A^L(d,j_L^*,\gamma_L^*)$ and some fixed and known vector $c^L(d,j_L^*,\gamma_L^*)$, $d \in \widehat{\mathcal{D}}$, $\hat{j}_L(d) = j_L^*$ and $\hat{\gamma}_L(d) = \gamma_L^*$ if and only if $A^L(d,j_L^*,\gamma_L^*)\hat{p} \leq c^L(d,j_L^*,\gamma_L^*)$, where $j_L^* \in \{1,...,J_L\}$ and γ_L^* is in the support of $\hat{\gamma}_L(d)$.*

2. For some finite-valued random vector $\hat{\gamma}_U(d)$, some fixed and known matrix $A^U(d,j_U^*,\gamma_U^*)$ and some fixed and known vector $c^U(d,j_U^*,\gamma_U^*)$, $d \in \widehat{\mathcal{D}}$, $\hat{j}_U(d) = j_U^*$ and $\hat{\gamma}_U(d) = \gamma_U^*$ if and only if $A^U(d,j_U^*,\gamma_U^*)\hat{p} \leq c^U(d,j_U^*,\gamma_U^*)$, where $j_U^* \in \{1,...,J_U\}$ and γ_U^* is in the support of $\hat{\gamma}_U(d)$.

Although not immediately obvious, this assumption holds in a variety of settings; see the examples below. In many cases, this assumption can be simplified because $d \in \widehat{\mathcal{D}}$ if and only if $A_{\mathcal{D}}\hat{p} \leq c_{\mathcal{D}}$ for some fixed and known matrix $A_{\mathcal{D}}$ and vector $c_{\mathcal{D}}$. For these cases, $\hat{\gamma}_L(d)$ and $\hat{\gamma}_U(d)$ can be vacuously set to fixed constants and

$$A^L(d,j,\gamma) = \begin{pmatrix} A_L(d,j) \\ A_{\mathcal{D}} \end{pmatrix}, \quad A^U(d,j,\gamma) = \begin{pmatrix} A_U(d,j) \\ A_{\mathcal{D}} \end{pmatrix}$$

11

and

$$c^L(d,j,\gamma) = \begin{pmatrix} c_L(d,j) \\ c_{\mathcal{D}} \end{pmatrix} \quad c^U(d,j,\gamma) = \begin{pmatrix} c_U(d,j) \\ c_{\mathcal{D}} \end{pmatrix}$$

for $A_L(d,j) = (\ell_{d,1} - \ell_{d,j},...,\ell_{d,J_L} - \ell_{d,j})'$, $A_U(d,j) = (u_{d,j} - u_{d,1},...,u_{d,j} - u_{d,J_L})'$, $c_L(d,j) = (\tilde{\ell}_{d,j} - \tilde{\ell}_{d,1},...,\tilde{\ell}_{d,j} - \tilde{\ell}_{d,J_L})'$ and $c_U(d,j) = (\tilde{u}_{d,1} - \tilde{u}_{d,j},...,\tilde{u}_{d,J_L} - \tilde{u}_{d,j})'$. A leading example of this special case is

$$\widehat{\mathcal{D}} = \left\{ d \in \{0,...,T\} : \widehat{U}(d) \geq \max_{d \in \{0,...,T\}} \widehat{L}(d) \right\},$$

where $\widehat{L}(d) \equiv \max_{j \in \{1,...,J_L\}} \{\tilde{\ell}_{d,j} + \ell_{d,j}\hat{p}\}$ and $\widehat{U}(d) \equiv \min_{j \in \{1,...,J_U\}} \{\tilde{u}_{d,j} + u_{d,j}\hat{p}\}$, since $d \in \widehat{\mathcal{D}}$ if and only if

$$(\ell_{d',j'} - u_{d,j})\hat{p} \leq \tilde{u}_{d,j} - \tilde{\ell}_{d',j'}$$

for all $d' \in \{0,...,T\}$, $j \in \{1,...,J_U\}$ and $j' \in \{1,...,J_L\}$.

We also note that Assumption 3.2 is compatible with the absence of data-dependent selection for which the researcher is interested in forming a CI for an identified interval $[L(d^*), U(d^*)]$ chosen by the researcher a priori. In these cases, $\widehat{\mathcal{D}} = \{d^*\}$ and $\hat{\gamma}_M(d^*)$ can be vacuously set to a fixed constant, $A^M(d^*,j,\gamma) = A_M(d^*,j)$ and $c^M(d^*,j,\gamma) = c_M(d^*,j)$ for $M = L,U$. Indeed, we examine an example of this special case when conducting a finite-sample power comparison in Section 7 below.

In general, less conditioning is more desirable in terms of the lengths of the CIs we propose. Although conditioning on the events $d \in \widehat{\mathcal{D}}$, $\hat{j}_L(d) = j_L^*$ and $\hat{j}_U(d) = j_U^*$ is necessary to construct our CIs (see Section 5 below), the researcher should therefore minimize the number of elements in $\hat{\gamma}_L(d)$ and $\hat{\gamma}_U(d)$ subject to satisfying Assumption 3.2, when constructing our CIs. In some cases it is necessary to condition on these additional random vectors in order to satisfy Assumption 3.2. But in many cases, such as the example given immediately above, additional conditioning random vectors are unnecessary and can be vacuously set to fixed constants.

We suppose that the sample of data is drawn from some unknown distribution $\mathbb{P} \in \mathcal{P}_n$. As an estimator for p, we assume that $\hat{p}$ is uniformly asymptotically normal under $\mathbb{P} \in \mathcal{P}_n$.

Assumption 3.3. *For the class of Lipschitz functions that are bounded in absolute value by one*

and have Lipschitz constant bounded by one, BL_1, there exist functions $p(\mathbb{P})$ and $\Sigma(\mathbb{P})$ such that for $\xi_{\mathbb{P}} \sim \mathcal{N}(0, \Sigma(\mathbb{P}))$ with

$$\lim_{n \to \infty} \sup_{\mathbb{P} \in \mathcal{P}_n} \sup_{f \in BL_1} \left| E_{\mathbb{P}}\left[f\left(\sqrt{n}(\hat{p} - p(\mathbb{P})) \right) \right] - E_{\mathbb{P}}[f(\xi_{\mathbb{P}})] \right| = 0.$$

The notation of this assumption makes explicit that the parameter p and the asymptotic variance Σ depend upon the unknown distribution of the data $\mathbb{P}$. It holds naturally for standard estimators $\hat{p}$ under random sampling or weak dependence in the presence of bounds of the moments and dependence of the underlying data.

Next, we assume that the asymptotic variance of $\hat{p}$ can be uniformly consistently estimated by an estimator $\widehat{\Sigma}$.

Assumption 3.4. *For all $\varepsilon > 0$, the estimator $\widehat{\Sigma}$ satisfies*

$$\lim_{n \to \infty} \sup_{\mathbb{P} \in \mathcal{P}_n} \mathbb{P}\left(\left\| \widehat{\Sigma} - \Sigma(\mathbb{P}) \right\| > \varepsilon \right) = 0.$$

This assumption is again naturally satisfied when using a standard sample analog estimator of Σ under random sampling or weak dependence in the presence of moment and dependence bounds.

In addition, we restrict the asymptotic variance of $\hat{p}$ to be positive definite.

Assumption 3.5. *For some finite $\bar{\lambda} > 0$, $1/\bar{\lambda} \leq \lambda_{\min}(\Sigma(\mathbb{P})) \leq \lambda_{\max}(\Sigma(\mathbb{P})) \leq \bar{\lambda}$ for all $\mathbb{P} \in \mathcal{P}_n$.*

This assumption is naturally satisfied, for example, when $\hat{p}$ is a standard sample analog estimator of reduced-form probabilities composing p that are non-redundant and bounded away from zero and one.

We impose one final assumption for our unconditional hybrid confidence intervals in order for the object of inferential interest to be well-defined unconditionally.

Assumption 3.6. $\widehat{\mathcal{D}} = \{\hat{d}\}$ *almost surely for a random variable $\hat{d}$ with support $\{0, ..., T\}$.*

In conjunction, Assumptions 3.2 and 3.6 hold naturally when the object of interest $\hat{d}$ is selected by uniquely maximizing a linear combination of the estimates of the bounds characterizing the

identified intervals and the additional conditioning vectors $\hat{\gamma}_L(d)$ and $\hat{\gamma}_U(d)$ are defined appropriately. Leading examples of this form of selection include when $\hat{d}$ corresponds to the largest estimated lower bound, upper bound or weighted average of lower and upper bounds.

Proposition 3.1. *Suppose $\widehat{\mathcal{D}} = \{\hat{d}\}$, where $\hat{d} = \text{argmax}_{d \in \{0,...,T\}} \{w_L \widehat{L}(d) + w_U \widehat{U}(d)\}$ is unique almost surely for some fixed known weights $w_L, w_U \geq 0$. Then Assumptions 3.2 and 3.6 are satisfied for*

1. *$\hat{\gamma}_L(d)$ equal to any fixed constant and $\hat{\gamma}_U(d) = \hat{\hat{j}}_L(d)$ when $w_U = 0$,*

2. *$\hat{\gamma}_L(d) = \hat{\gamma}_U(d) = (\hat{\hat{j}}_U(0),...,\hat{\hat{j}}_U(T))'$ when $w_L = 0$,*

3. *$\hat{\gamma}_L(d) = \hat{\gamma}_U(d) = (\hat{\hat{j}}_L(0),...,\hat{\hat{j}}_L(T),\hat{\hat{j}}_U(0),...,\hat{\hat{j}}_U(T))'$ when $w_L, w_U \neq 0$.*

As this proposition makes clear, the additional conditioning vectors needed for Assumption 3.6 to hold depend upon the particular form of selection rule used by the researcher. For example, when $\hat{d}$ is chosen to maximize the estimated lower bound of the identified set $\widehat{L}(d)$, one must condition not only on the realized value of $\hat{\hat{j}}_U(\hat{d})$ when forming a probabilistic upper bound for $U(\hat{d})$ but also $\hat{\hat{j}}_L(\hat{d})$. On the other hand, the formation of either a probabilistic lower bound for $L(\hat{d})$ or upper bound for $U(\hat{d})$ when $\hat{d}$ is chosen to maximize the estimated upper bound of the identified set $\widehat{U}(d)$ requires conditioning on the entire vector $(\hat{\hat{j}}_U(0),...,\hat{\hat{j}}_U(T))'$.

Although intuitively appealing, the treatment choice rules of the form described in Proposition 3.1 can be sub-optimal from a statistical decision-theoretic point of view (see, e.g., Manski, 2021, 2023 and Christensen et al., 2023). In Section 4.6, we show how proper definition of $\hat{\gamma}_L(d)$ and $\hat{\gamma}_U(d)$ satisfies Assumptions 3.2 and 3.6 in the context of optimal selection rules.

4 Examples

In this section, we show that the proposed inference method is applicable to various examples for which parameters are interval-identified. In particular, we show that Assumptions 3.1–3.3 are satisfied in these examples.

4.1 Revisiting Manski Bounds with a Continuous Outcome

We revisit the example with Manski bounds in Section 2, now allowing for a continuous outcome with bounded support. Let $Y \in [y^l, y^u]$ be continuously distributed and suppose $E[Y(d)|Z] = E[Y(d)]$ for $d \in \{0,1\}$. Note that

$$L(d) \equiv \max_{z \in \{0,1\}} [E[Y|D=d, Z=z]P(D=d|Z=z) + (1-P(D=d|Z=z))y^l],$$

$$U(d) \equiv \min_{z \in \{0,1\}} [E[Y|D=d, Z=z]P(D=d|Z=z) + (1-P(D=d|Z=z))y^u],$$

and $L(1,0)$ and $U(1,0)$ are Manski's sharp bounds on the ATE for

$$L(\tilde{d}, d) \equiv L(\tilde{d}) - U(d),$$

$$U(\tilde{d}, d) \equiv U(\tilde{d}) - L(d).$$

We can define the identified set $\mathcal{D}^* \subseteq \mathcal{D}$ of optimal treatments as

$$\mathcal{D}^* \equiv \{d \in \{0,1\} : L(\tilde{d}, d) \leq 0, \forall \tilde{d} \in \{0,1\}\} = \{d \in \{0,1\} : \max_{\tilde{d} \in \{0,1\}} L(\tilde{d}) \leq U(d)\}.$$

Then, Assumption 3.1 holds with

$$p = \begin{pmatrix} E(Y|D=0, Z=1)P(D=0|Z=1) \\ E(Y|D=0, Z=0)P(D=0|Z=0) \\ E(Y|D=1, Z=1)P(D=1|Z=1) \\ E(Y|D=1, Z=0)P(D=1|Z=0) \\ P(D=0|Z=1) \\ P(D=0|Z=0) \end{pmatrix}$$

and

$$\tilde{\ell}_{0,j} = y^l \text{ for } j \in \{0,1\},$$

$$\tilde{\ell}_{1,j} = 0 \text{ for } j \in \{0,1\},$$

$$\ell_{0,0} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & -y^l \end{pmatrix},$$

$$\ell_{0,1} = \begin{pmatrix} 1 & 0 & 0 & 0 & -y^l & 0 \end{pmatrix},$$

$$\ell_{1,0} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & y^l \end{pmatrix},$$

$$\ell_{1,1} = \begin{pmatrix} 0 & 0 & 1 & 0 & y^l & 0 \end{pmatrix}$$

and symmetrically for $\tilde{u}_{d,j}$ and $u_{d,j}$. For $\widehat{L}(d)$ and $\widehat{U}(d)$ being the sample counterparts of $L(d)$ and $U(d)$ upon replacing p with $\hat{p}$, Assumption 3.2 holds by the argument in the paragraph after Assumption 3.2 since

$$\widehat{\mathcal{D}} = \{d \in \{0,1\} : \max_{\tilde{d} \in \{0,1\}} \widehat{L}(\tilde{d}) \leq \widehat{U}(d)\}.$$

For Assumption 3.3, let $[y^l, y^u] = [0,1]$ for simplicity. Note that

$$E(Y|D=d,Z=z)P(D=d|Z=z) = P(D=d|Z=z) - \int_0^1 P(Y \leq y, D=d|Z=z)dy$$

$$= P(D=d|Z=z) - E\left[\int_0^1 1(Y \leq y, D=d)dy \middle| Z=z\right],$$

where the first equality uses integration by parts. Therefore, we can estimate the elements of p by sample means, forming $\hat{p}$.

4.2 Bounds Derived from Linear Programming

In more complex settings, calculating analytical bounds on $W(d)$ or $W(d) - W(\tilde{d})$ may be cumbersome. This is especially true when the researcher wants to incorporate additional identifying assumptions. In this situation, the computational approach using linear programming can be

useful (Mogstad et al., 2018; Han and Yang, 2024).

To incorporate many complicated settings, suppose that $W(d) = A_d q$ and $p = Bq$ for some known row vector A_d and matrix B, an unknown vector q in a simplex $\mathcal{Q}$, and a vector p that is estimable from data. Typically q is a vector of probabilities of a latent variable that governs the DGP; see Balke and Pearl (1997, 2011), Mogstad et al. (2018), Han (2024) and Han and Yang (2024). The linearity in this assumption is usually implied by the nature of a particular problem (e.g., discreteness). Then we have

$$\begin{aligned} L(d) &= \min_{q \in \mathcal{Q}} A_d q, \\ U(d) &= \max_{q \in \mathcal{Q}} A_d q, \end{aligned} \qquad s.t. \quad Bq = p \qquad (4.1)$$

and ATE bounds for a change from treatment d to treatment $\tilde{d}$

$$\begin{aligned} L(\tilde{d},d) &= \min_{q \in \mathcal{Q}} (A_{\tilde{d}} - A_d) q, \\ U(\tilde{d},d) &= \max_{q \in \mathcal{Q}} (A_{\tilde{d}} - A_d) q, \end{aligned} \qquad s.t. \quad Bq = p. \qquad (4.2)$$

Note that $L(\tilde{d},d) \neq L(\tilde{d}) - U(d)$ in general because the q that solves (4.1) for $L(\tilde{d})$ and $U(d)$ may be different (and similarly for $U(\tilde{d},d)$). As before, the identified set of optimal treatments here is characterized as $\mathcal{D}^* \equiv \{d : L(\tilde{d},d) \leq 0, \forall \tilde{d} \neq d\}$.

An example of this setting is as follows. Let (Y,D,Z) be a vector of a binary outcome, treatment and instrument and let p be a vector with entries $p(y,d|z) \equiv P(Y=y,D=d|Z=z)$ across $(y,d,z) \in \{0,1\}^3$.[2] Suppose $W(d) = E[Y(d)]$ for $d \in \{0,1\}$. Then, we can define the response type $\varepsilon \equiv (Y(1),Y(0),D(1),D(0))$ with a realized value $e \equiv (y(1),y(0),d(1),d(0))$, where $Y(d)$ denotes the potential outcome under treatment d and $D(z)$ denotes the potential treatment under instrument value z. Let $q(e) \equiv P(\varepsilon = e)$ be the latent distribution. Then

$$W(d) = \Pr[Y(d) = 1] = \sum_{e:y(d)=1} q(e) \equiv A_d q,$$

---

[2]See Han and Yang (2024) for the use of linear programming with continuous Y.

where q is the vector of $q(e)$'s and A_d is an appropriate selector (a row vector).

Assume that $(Y(d),D(z))$ is independent of Z for $d,z \in \{0,1\}$. The data distribution p is related to the latent distribution by

$$\Pr[Y=1,D=d|Z=z]=\Pr[Y(d)=1,D(z)=d]=\sum_{e:y(d)=1,d(z)=d} q(e) \equiv B_{d,z}q,$$

where the first equality follows by the independence assumption., q is a vector of $q(e)$'s and $B_{d,z}$ is an appropriate selector (a row vector). Now define

$$B \equiv \begin{bmatrix} B_{1,1} \\ B_{0,1} \\ B_{1,0} \\ \vdots \end{bmatrix}, \qquad p \equiv \begin{bmatrix} p(1,1|1) \\ p(1,0|1) \\ p(1,1|0) \\ p(1,0|0) \\ \vdots \end{bmatrix}$$

so that all of the constraints relating the data distribution to the latent distribution can be expressed as $Bq=p$.

To verify Assumption 3.1, it is helpful to invoke strong duality for the primal problems (4.1) (under regularity conditions) and write the following dual problems:

$$L(d)=\max_{\lambda}-\tilde{p}'\lambda, \quad s.t. \quad \tilde{B}'\lambda \geq -A_d',$$

$$U(d)=\min_{\lambda}\tilde{p}'\lambda, \quad s.t. \quad \tilde{B}'\lambda \geq A_d',$$

where $\tilde{B} \equiv \begin{bmatrix} B \\ \mathbf{1}' \end{bmatrix}$ is a $(d_p+1) \times d_q$ matrix with $\mathbf{1}$ being a $d_q \times 1$ vector of ones, and $\tilde{p} \equiv \begin{bmatrix} p \\ 1 \end{bmatrix}$ is a $(d_p+1) \times 1$ vector. By using a vertex enumeration algorithm (e.g., Avis and Fukuda (1992)), one can find all (or a relevant subset) of vertices of the polyhedra $\{\lambda:\tilde{B}'\lambda \geq -A_d'\}$ and $\{\lambda:\tilde{B}'\lambda \geq A_d'\}$. Let $\Lambda_{L,d} \equiv \{\lambda_1,...,\lambda_{J_{L,d}}\}$ and $\Lambda_{U,d} \equiv \{\lambda_1,...,\lambda_{J_{U,d}}\}$ be the sets that collect such vertices, respectively. Then, it is easy to see that $L(d)=\max_{\lambda \in \Lambda_{L,d}}-\tilde{p}'\lambda$ and $U(d)=\min_{\lambda \in \Lambda_{U,d}}\tilde{p}'\lambda$, and thus Assumption 3.1 holds.

To verify Assumption 3.2, we use the dual problems to (4.2):

$$L(\tilde{d},d)=\max_{\lambda}-\tilde{p}'\lambda, \quad s.t. \quad \tilde{B}'\lambda\geq-\Delta'_{\tilde{d},d},$$

$$U(\tilde{d},d)=\min_{\lambda}\tilde{p}'\lambda, \quad s.t. \quad \tilde{B}'\lambda\geq\Delta'_{\tilde{d},d},$$

where $\Delta_{\tilde{d},d} \equiv A_{\tilde{d}} - A_d$. Analogous to the vertex enumeration argument above, let $\Lambda_{L,\tilde{d},d} \equiv \{\lambda_1,...,\lambda_{J_{L,\tilde{d},d}}\}$ and $\Lambda_{U,\tilde{d},d} \equiv \{\lambda_1,...,\lambda_{J_{U,\tilde{d},d}}\}$ be the sets that collect all (or a relevant subset) of vertices of the polyhedra $\{\lambda : \tilde{B}'\lambda \geq -\Delta'_{\tilde{d},d}\}$ and $\{\lambda : \tilde{B}'\lambda \geq \Delta'_{\tilde{d},d}\}$, respectively. Then, $L(\tilde{d},d) = \max_{\lambda\in\Lambda_{L,\tilde{d},d}} -\tilde{p}'\lambda$ and $U(\tilde{d},d) = \min_{\lambda\in\Lambda_{U,\tilde{d},d}} \tilde{p}'\lambda$. Let $\widehat{\mathcal{D}} = \{d : \widehat{L}(\tilde{d},d) \leq 0, \forall \tilde{d} \neq d\}$ where $\widehat{L}(\tilde{d},d)$ is the sample counterpart of $L(\tilde{d},d)$ with $\widehat{\tilde{p}}\equiv\begin{bmatrix}\hat{p}\\1\end{bmatrix}$ replacing $\tilde{p}\equiv\begin{bmatrix}p\\1\end{bmatrix}$. Partition λ as $\lambda=(\lambda^{1\prime},\lambda^0)'$ where λ^0 is the last element of λ. Note that $d\in\widehat{\mathcal{D}}$ if and only if

$$\max_{\lambda\in\tilde{\Lambda}_{L,d}} -(\hat{p}'\lambda^1+\lambda^0)\leq0,$$

where $\tilde{\Lambda}_{L,d}=\bigcup_{\tilde{d}\neq d}\Lambda_{L,\tilde{d},d}$. Also let $\hat{\lambda}$ be such that $-\widehat{\tilde{p}}'\hat{\lambda}=\max_{\lambda\in\tilde{\Lambda}_{L,d}}-\widehat{\tilde{p}}'\lambda$. Then, $\hat{\lambda}=\lambda_{j_L^*}$ if and only if

$$\hat{p}'\lambda_{j_L^*}^1+\lambda_{j_L^*}^0-(\hat{p}'\lambda^1+\lambda^0)\leq0 \quad \forall\lambda\in\tilde{\Lambda}_{L,d}\backslash\{\lambda_{j_L^*}\}$$

so that Assumption 3.2 holds.

Finally, $\hat{p}$ is again equal to a vector of sample means so that Assumption 3.3 is satisfied if $\hat{p}$ is calculated using the random sample $\{Y_i,D_i,Z_i\}_{i=1}^n$.

4.3 Empirical Welfare Maximization with Observational Data

Consider allocating a binary treatment based on observed covariates $X\in\mathcal{X}$. A treatment allocation rule can be defined as a function $\delta:\mathcal{X}\to\{0,1\}$ in a class of rules $\mathcal{D}$. Consider the utilitarian welfare of deploying δ relative to treating no one (Manski (2009)): $W(\delta)=E[Y(\delta(X))-Y(0)]$.

The optimal allocation δ^* satisfies

$$\delta^* \in \underset{\delta \in \mathcal{D}}{\operatorname{argmax}} W(\delta).$$

Note that

$$E[Y(\delta(X)) - Y(0)] = E[\delta(X)\Delta(X)],$$

where $\Delta(X) \equiv E[Y(1) - Y(0)|X]$. This problem is considered in Kitagawa and Tetenov (2018) and Athey and Wager (2021), among others. When only observational data for (Y, D, X) are available with D being endogenous, $W(\delta)$ is only partially identified unless strong treatment effect homogeneity is assumed. This problem has been studied in Kallus and Zhou (2021), Pu and Zhang (2021), D'Adamo (2021) and Byambadalai (2021), among others. Using instrumental variables, one can consider bounds on the conditional ATE based on conditional versions of the bounds considered in Sections 4.1 and 4.2 (i.e., Manski's bounds or bounds produced by linear programming).

In particular, assume that $(Y(d), D(z))$ is independent of Z given X. Let $L(X)$ and $U(X)$ be conditional Manski bounds on $\Delta(X)$. Then, bounds on $W(\delta)$ can be characterized as

$$L(\delta) \equiv E[\delta(X)L(X)],$$
$$U(\delta) \equiv E[\delta(X)U(X)].$$

Similarly, bounds on $W(\tilde{\delta}) - W(\delta) = E[(\tilde{\delta}(X) - \delta(X))\Delta(X)]$ can be characterized as

$$L(\tilde{\delta}, \delta) \equiv E[(\tilde{\delta}(X) - \delta(X))L(X)], \qquad (4.3)$$
$$U(\tilde{\delta}, \delta) \equiv E[(\tilde{\delta}(X) - \delta(X))U(X)].$$

Note that $L(\tilde{\delta}, \delta) \neq L(\tilde{\delta}) - U(\delta)$ in general (and similarly for $U(\tilde{\delta}, \delta)$).

Suppose $\mathcal{X}$ is finite and $\mathcal{X} = \{x_1, ..., x_K\}$ where x_k can be a vector and K can potentially be large. For simplicity of exposition, suppose $\mathcal{X} = \{0, 1, 2\}$. Then $\mathcal{D} = \{\delta_1, ..., \delta_8\}$ where each δ_j

corresponds to a mapping type from $\{0,1,2\}$ to $\{0,1\}$. To verify Assumptions 3.1 and 3.2, we proceed as follows. For given $x \in \mathcal{X}$, by arguments analogous to those in Sections 4.1 and 4.2, bounds L_x and U_x on $\Delta(x)$ satisfy, for some scalars $\tilde{\ell}_j$ and $\tilde{u}_j$ and row vectors ℓ_j and u_j,

$$L_x = \max_{j \in \{1,\dots,J_L\}} \{\tilde{\ell}_j + \ell_j p(x)\},$$

$$U_x = \min_{j \in \{1,\dots,J_U\}} \{\tilde{u}_j + u_j p(x)\},$$

where $p(x)$ is the vector of $p(y,d|z,x)$'s across (y,d,z). Then, by Jensen's inequality, for each $\delta \in \mathcal{D}$,

$$L(\delta) \geq \tilde{L}(\delta) \equiv \max_{j \in \{1,\dots,J_L\}} \{\tilde{\ell}_j E[\delta(X)] + \ell_j E[\delta(X)p(X)]\},$$

$$U(\delta) \leq \tilde{U}(\delta) \equiv \min_{j \in \{1,\dots,J_U\}} \{\tilde{u}_j E[\delta(X)] + u_j E[\delta(X)p(X)]\}.$$

Note that $\tilde{L}(\delta)$ and $\tilde{U}(\delta)$ are non-sharp bounds; for calculation of sharp bounds, see Section 4.4. We can verify Assumption 3.1 with $\tilde{L}(\delta)$ and $\tilde{U}(\delta)$ by defining

$$p = \begin{pmatrix} E[\delta_1(X)] \\ E[\delta_1(X)p(X)] \\ \vdots \\ E[\delta_8(X)] \\ E[\delta_8(X)p(X)] \end{pmatrix}$$

and, for $\delta = \delta_1$ as an example, by using

$$\ell_{\delta_1,j} = (\begin{array}{ccccc} \tilde{\ell}_j & \ell_j & 0 & \cdots & 0 \end{array}).$$

Similarly, we can verify Assumptions 3.2 and 3.3 by estimating $p(X)$ and $E[\delta(X)]$ with sample means and $E[\delta(X)p(X)]$ with $\frac{1}{n}\sum_i^n \delta(X_i)\hat{p}(X_i)$. If the data $\{Y_i,D_i,Z_i,X_i\}_{i=1}^n$ form a random sample, and $\widehat{\mathcal{D}} = \{\delta \in \mathcal{D} : \widehat{L}(\tilde{\delta},\delta) \leq 0 \forall \tilde{\delta} \neq \delta\}$ for $\widehat{L}(\tilde{\delta},\delta)$ defined the same as $L(\tilde{\delta},\delta)$ in (4.3) after substituting $\hat{p}$ for p, Assumptions 3.2 and 3.3 hold.

21

This framework can be generalized to settings where $W(\delta)$ is partially identified, not necessarily due to treatment endogeneity but because $W(\delta)$ is a non-utilitarian welfare defined as a functional of the joint distribution of potential outcomes (e.g., Cui and Han, 2024): $W(\delta) = f(F_{Y(1),Y(0)|X})$ where f is some functional and $F_{Y(1),Y(0)|X}$ is the joint distribution of $(Y(1),Y(0))$ conditional on X.

4.4 Empirical Welfare Maximization via Linear Programming

Continuing from Section 4.3, we show how sharp bounds on $W(\delta)$ can be computed using linear programming. This can be done by extending the example in Section 4.2 with binary Y. Again, let $\mathcal{X} = \{x_1,...,x_K\}$. Let $q(e|x) \equiv P(\varepsilon = e | X = x)$. In analogy,

$$E[Y(d)|X=x] = \Pr[Y(d)=1|X=x] = \sum_{e:y(d)=1} q(e|x) \equiv A_d q(x),$$

where $q(x)$ is a vector with entries $q(e|x)$ across e, and

$$\Pr[Y=1, D=d | Z=z, X=x] = \Pr[Y(d)=1, D(z)=d | X=x]$$

$$= \sum_{e:y(d)=1, d(z)=d} \Pr[\varepsilon=e | X=x] \equiv B_{d,z} q(x),$$

where the first equality holds by the independence assumption in the previous section. Then the constraint for each $x \in \mathcal{X}$ becomes

$$Bq(x) = p(x).$$

Now we can construct a linear program for welfare:

$$W(\delta) = E[\delta(X)\Delta(X)] = \sum_{x_k \in \mathcal{X}} p(x_k)\delta(x_k)\Delta(x_k)$$

$$= \sum_{x_k \in \mathcal{X}} p(x_k)\delta(x_k)(A_1 - A_0)q(x_k).$$

Therefore $W(\delta)$ satisfies the structure of Section 4.2 and by analogous arguments, Assumptions 3.1–3.3 hold.

4.5 Bounds for Dynamic Treatment Effects

Consider binary Y_t and D_t for $t = 1,...,T$. Let $Y \equiv (Y_1,...,Y_T)$ and $D \equiv (D_1,...,D_T)$. Suppose that we are equipped with a sequence of binary instruments $Z \equiv (Z_{t_1},...,Z_{t_K})$, which is a subvector of $(Z_1,...,Z_T)$. For $t = 1,...,T$, let $Y_t(d_1,...,d_t)$ be the potential outcome at t and $Y(d) \equiv (Y_1(d_1),...,Y_T(d_1,...,d_T))$. We assume that the instruments Z are independent of the potential outcomes $Y(d)$.

Let $T=2$. Then $Y \equiv (Y_1,Y_2)$ and $D \equiv (D_1,D_2)$. We are interested in the optimal policy d^* that satisfies

$$d^* \in \underset{d \in \mathcal{D}}{\mathrm{argmax}} W(d),$$

where $\mathcal{D} \equiv \{(1,1),(1,0),(0,1),(0,0)\}$. The sign of the welfare difference, $W(d)-W(\tilde{d})$ for $d,\tilde{d} \in \mathcal{D}$, is useful for establishing the ordering of $W(d)$ with respect to d and thus to identifying d^*. However, without additional identifying assumptions, we can only establish a partial ordering of $W(d)$ based on the bounds on the welfare difference. This will produce the identified set $\mathcal{D}^*$ for d^*.

An example of the welfare is $W(d) \equiv E[Y_2(d)]$, namely, the average potential terminal outcome. The bounds on welfare $W(d)$ are

$$L(d) \equiv \max_z L(d;z), \tag{4.4}$$

$$U(d) \equiv \min_z U(d;z), \tag{4.5}$$

where

$$L(d;z) \equiv P(Y_2=1,D=d|Z=z),$$

$$U(d;z) \equiv P(Y_2=1,D=d|Z=z)+\sum_{d'\neq d}P(D=d'|Z=z),$$

23

which have forms analogous to those in the static case. Define the dynamic ATE in the terminal period for a change in treatment from d to $\tilde{d}$ as

$$W(\tilde{d})-W(d)=E[Y_2(\tilde{d})-Y_2(d)]=P(Y_2(\tilde{d})=1)-P(Y_2(d)=1).$$

Then the bounds on the dynamic ATE are as follows:

$$L(\tilde{d},d)\equiv L(\tilde{d})-U(d), \tag{4.6}$$

$$U(\tilde{d},d)\equiv U(\tilde{d})-L(d). \tag{4.7}$$

Another example of welfare is the joint distribution $W(d)\equiv P(Y(d)=(1,1))$ where $Y(d)\equiv (Y_1(d_1),Y_2(d))$. The bounds on $W(d)$ in this case are $L(d)\equiv\max_z L(d;z)$ and $U(d)\equiv\min_z U(d;z)$ where

$$L(d;z)\equiv P(Y_2=1,D=d|Z=z),$$

$$U(d;z)\equiv P(Y_2=1,D=d|Z=z)+P(Y_1=1,D_1=d_1,D_2=d_2'|Z=z)$$

$$+P(D_1=d_1',D_2=d_2|Z=z)+P(D_1=d_1',D_2=d_2'|Z=z).$$

Consider the effect of treatment on the joint distribution, for example,

$$W(1,1)-W(1,0)=P(Y(1,1)=(1,1))-P(Y(1,0)=(1,1)).$$

Then, with $\tilde{d}=(1,1)$ and $d=(1,0)$, the bounds on this parameter are

$$
\begin{aligned}
L(\tilde{d},d) \equiv{}& \max_z P(Y_2=1,D=(1,1)|Z=z) \\
& -\min_z[P(Y_2=1,D=(1,0)|Z=z)+P(Y_1=1,D=(1,1)|Z=z) \\
& +P(D=(0,1)|Z=z)+P(D=(0,0)|Z=z)], \\
U(\tilde{d},d) \equiv{}& \min_z[P(Y_2=1,D=(1,1)|Z=z)+P(Y_1=1,D=(1,0)|Z=z) \\
& +P(D=(0,1)|Z=z)+P(D=(0,0)|Z=z)] \\
& -\max_z P(Y_2=1,D=(1,0)|Z=z).
\end{aligned}
$$

In these examples, the identified set $\mathcal{D}^*$ can be characterized as a set of maximal elements:

$$
\begin{aligned}
\mathcal{D}^* &= \{d : \nexists \tilde{d}\neq d \text{ such that } L(\tilde{d},d)>0\} \\
&= \{d : L(\tilde{d},d)\leq 0, \forall \tilde{d}\neq d\}.
\end{aligned}
$$

These examples are special cases of the model in Han (2024).[3]

In both cases, it is easy to see that $L(d)$ and $U(d)$ satisfy Assumption 3.1 with p being the vector of probabilities $p(y,d|z)\equiv P(Y=y,D=d|Z=z)$. To verify Assumption 3.2, let $\widehat{L}(\tilde{d},d)$ be the estimator of $L(\tilde{d},d)$ where the sample frequency replaces the population probability. Then $\widehat{\mathcal{D}}=\{d:\widehat{L}(\tilde{d},d)\leq 0,\forall \tilde{d}\neq d\}$.

Continuing with the first example, let $Z=Z_1\in\{0,1\}$, that is, the researcher is only equipped with a binary instrument in the first period and no instrument in the second period. We focus on inference for $L(d)$ for $d\in\widehat{\mathcal{D}}$. Let $\widehat{z}(d)\in\mathrm{argmax}_{z\in\{0,1\}}\widehat{L}(d;z)$ so that $\widehat{L}(d)=\widehat{L}(d;\widehat{z}(d))$. Then we can write the data-dependent event that d is an element of $\widehat{\mathcal{D}}$, $\widehat{L}(d)=\widehat{L}(d;z^*)$ as a polyhedron

$$
\{d\in\widehat{\mathcal{D}},\widehat{z}(d)=z^*\}=\{\widehat{L}(\tilde{d},d)\leq 0\forall \tilde{d}\neq d,\widehat{z}(d)=z^*\}=\{A^L\hat{p}\leq 0\}
$$

---

[3] See also Han and Lee (2024) for other examples of dynamic causal parameters that can be used to define optimal treatments.

for some matrix A^L, where $\hat{p}$ is the vector of probabilities $\hat{p}(y,d|z)$, so that Assumption 3.2 holds. This is due to the forms of $L(d;z)$ and $U(d;z)$ above and $\widehat{L}(\tilde{d},d) \leq 0 \ \forall \tilde{d} \neq d$ if and only if

$$\widehat{L}(\tilde{d};z) \leq \widehat{U}(d;z') \quad \forall \tilde{d} \neq d, \forall z, z'$$

and, for example, $\widehat{z}(d) = 1$ if and only if

$$\widehat{L}(d;0) - \widehat{L}(d;1) \leq 0.$$

A similar formulation follows for the second example. In fact, this approach applies to a general parameter $W(d)$ with bounds that are minimum and maximums of linear combinations of $p(y,d|z)$'s, such as parameters that have the following form:

$$W(d) \equiv f(q_d)$$

for some linear functional f, where $q_d(y) \equiv P(Y(d) = y)$.

Finally, Assumption 3.3 is satisfied in both examples when the data $\{Y_i, D_i, Z_i\}_{i=1}^n$ form a random sample since the entries of $\hat{p}$ are sample means.

This framework can be further generalized to incorporate treatment choices adaptive to covariates or past outcomes as in Han (2024), analogous to Section 4.3. Sometimes, this generalization prevents the researcher from deriving analytical bounds, in which case the linear programming approach can be used.

4.6 Optimal Treatment Assignment with Interval-Identified ATE

In recent work, Christensen et al. (2023) note that "plug-in" rules for determining treatment choice can be sub-optimal when ATEs are not point-identified since the bounds on the ATE are not smooth functions of the reduced-form parameter p. Using an optimality criterion that minimizes maximum regret over the identified set for the ATE, conditional on p, they advocate bootstrap and quasi-Bayesian methods for optimal treatment choice. More specifically, they consider settings

for which the ATE of a treatment is identified via intersection bounds:

$$b_L(p) \equiv \max_{1 \leq k \leq K_L} \{\tilde{\mathfrak{l}}_k + \mathfrak{l}_k p\} \leq ATE \leq \min_{1 \leq k \leq K_U} \{\tilde{\mathfrak{u}}_k + \mathfrak{u}_k p\} \equiv b_U(p)$$

for some fixed and known K_L, K_U, $\tilde{\mathfrak{l}}_1, ..., \tilde{\mathfrak{l}}_{K_L}$, $\tilde{\mathfrak{u}}_1, ..., \tilde{\mathfrak{u}}_{K_U}$, $\mathfrak{l}_1, ..., \mathfrak{l}_{K_L}$ and $\mathfrak{u}_1, ..., \mathfrak{u}_{K_U}$.[4]

Christensen et al. (2023) advocate a quasi-Bayesian implementation of their optimal treatment choice rule taking the form

$$\hat{d} = \mathbf{1}\left(\frac{1}{m} \sum_{i=1}^{m} [\max\{b_U(\hat{p} + \varepsilon_i), 0\} + \min\{b_L(\hat{p} + \varepsilon_i), 0\}] \geq 0\right), \tag{4.8}$$

for some large m, where $\varepsilon_1, ..., \varepsilon_m \overset{i.i.d.}{\sim} \mathcal{N}(0, \widehat{\Sigma})$ are independent of $\hat{p}$. As the following proposition shows, this form of $\hat{d}$ satisfies Assumption 3.2 when $\hat{\gamma}_L(d) = \hat{\gamma}_U(d)$ are specified properly.

Proposition 4.1. *Suppose $\widehat{\mathcal{D}} = \{\hat{d}\}$, where $\hat{d}$ is defined by (4.8). Then Assumptions 3.2 and 3.6 are satisfied for*

$$\hat{\gamma}_L(d) = \hat{\gamma}_U(d) = (\varepsilon_1', ..., \varepsilon_m', \underline{k}_1, ..., \underline{k}_m, \bar{k}_1, ..., \bar{k}_m, s_1^{\mathfrak{l}}, ..., s_m^{\mathfrak{l}}, s_1^{\mathfrak{u}}, ..., s_m^{\mathfrak{u}})',$$

where $\underline{k}_i = \mathrm{argmin}_{1 \leq k \leq K_U} \{\tilde{\mathfrak{u}}_k + \mathfrak{u}_k(\hat{p} + \varepsilon_i)\}$, $\bar{k}_i = \mathrm{argmax}_{1 \leq k \leq K_U} \{\tilde{\mathfrak{l}}_k + \mathfrak{l}_k(\hat{p} + \varepsilon_i)\}$, $s_i^{\mathfrak{l}} = \mathrm{sign}(\tilde{\mathfrak{l}}_{\bar{k}_i} + \mathfrak{l}_{\bar{k}_i}(\hat{p} + \varepsilon_i))$ and $s_i^{\mathfrak{u}} = \mathrm{sign}(\tilde{\mathfrak{u}}_{\underline{k}_i} + \mathfrak{u}_{\underline{k}_i}(\hat{p} + \varepsilon_i))$ for $i = 1, ..., m$.

5 Conditional Confidence Intervals

We now generalize the conditional CI construction described in Section 2.2 to apply in the general framework of Section 3, covering all example applications discussed above. As in Section 2.2, we are interested in forming probabilistic lower and upper bounds $\widehat{L}(\hat{d})_\alpha^C$ and $\widehat{U}(\hat{d})_{1-\alpha}^C$ that satisfy (2.2) and (2.5) for all $d \in \{0, ..., T\}$ as endpoints in the formation of a conditionally valid CI because the researcher's interest in inference on option d only arises when it is a member of the estimated set $\widehat{\mathcal{D}}$.

---

[4]Although Christensen et al. (2023) do not write the form of their bounds as they are written here, the representation here is equivalent to the one in that paper upon proper definition of p since the elements in the intersection bounds are smooth functions of a reduced-form parameter.

To begin, we characterize the conditional distributions of $\widehat{L}(d)$ given $d \in \widehat{\mathcal{D}}$, $\hat{\jmath}_L(d) = j_L^*$ and $\hat{\gamma}_L(d) = \gamma_L^*$ and $\widehat{U}(d)$ given $d \in \widehat{\mathcal{D}}$, $\hat{\jmath}_U(d) = j_U^*$ and $\hat{\gamma}_U(d) = \gamma_U^*$. These conditional distributions depend upon the nuisance parameter p. As a first step, form sufficient statistics for p that are asymptotically independent of $\widehat{L}(d)$ given $\hat{\jmath}_L(d) = j_L^*$ and $\widehat{U}(d)$ given $\hat{\jmath}_U(d) = j_U^*$:

$$\widehat{Z}_L(d,j_L^*) = \sqrt{n}\hat{p} - \hat{b}_L(d,j_L^*)\sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p})$$

and

$$\widehat{Z}_U(d,j_U^*) = \sqrt{n}\hat{p} - \hat{b}_U(d,j_U^*)\sqrt{n}(\tilde{u}_{d,j_U^*} + u_{d,j_U^*}\hat{p})$$

with

$$\hat{b}_L(d,j_L^*) = \widehat{\Sigma}\ell'_{d,j_L^*}\left(\ell_{d,j_L^*}\widehat{\Sigma}\ell'_{d,j_L^*}\right)^{-1} \quad \text{and} \quad \hat{b}_U(d,j_U^*) = \widehat{\Sigma}u'_{d,j_U^*}\left(u_{d,j_U^*}\widehat{\Sigma}u'_{d,j_U^*}\right)^{-1}.$$

Under Assumption 3.2, Lemma 1 of McCloskey (2024) implies

$$\sqrt{n}\widehat{L}(d)\Big|\left\{d \in \widehat{\mathcal{D}}, \hat{\jmath}_L(d) = j_L^*, \hat{\gamma}_L(d) = \gamma_L^*\right\}$$
$$\sim \sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p})\Big|\left\{\widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d,j_L^*),d,j_L^*,\gamma_L^*\right) \leq \sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p}) \leq \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L(d,j_L^*),d,j_L^*,\gamma_L^*\right),\right.$$
$$\left.\widehat{\mathcal{V}}_L^0\left(\widehat{Z}_L(d,j_L^*),d,j_L^*,\gamma_L^*\right) \geq 0\right\}$$

and

$$\sqrt{n}\widehat{U}(d)\Big|\left\{d \in \widehat{\mathcal{D}}, \hat{\jmath}_U(d) = j_U^*, \hat{\gamma}_U(d) = \gamma_U^*\right\}$$
$$\sim \sqrt{n}(\tilde{u}_{d,j_U^*} + u_{d,j_U^*}\hat{p})\left\{\widehat{\mathcal{V}}_U^-\left(\widehat{Z}_U(d,j_U^*),d,j_U^*,\gamma_U^*\right) \leq \sqrt{n}(\tilde{u}_{d,j_U^*} + u_{d,j_U^*}\hat{p}) \leq \widehat{\mathcal{V}}_U^+\left(\widehat{Z}_U(d,j_U^*),d,j_U^*,\gamma_U^*\right),\right.$$
$$\left.\widehat{\mathcal{V}}_U^0\left(\widehat{Z}_U(d,j_U^*),d,j_U^*,\gamma_U^*\right) \geq 0\right\}$$

with

$$\widehat{\mathcal{V}}_M^-(z,d,j,\gamma) = \max_{k:(A^M(d,j,\gamma)\hat{b}_M(d,j))_k<0} \frac{\sqrt{n}(c^M(d,j,\gamma))_k - (A^M(d,j,\gamma)z)_k}{(A^M(d,j,\gamma)\hat{b}_M(d,j))_k},$$

$$\widehat{\mathcal{V}}_M^+(z,d,j,\gamma) = \min_{k:(A^M(d,j,\gamma)\hat{b}_M(d,j))_k>0} \frac{\sqrt{n}(c^M(d,j,\gamma))_k - (A^M(d,j,\gamma)z)_k}{(A^M(d,j,\gamma)\hat{b}_M(d,j))_k},$$

$$\widehat{\mathcal{V}}_M^0(z,d,j,\gamma) = \min_{k:(A^M(d,j,\gamma)\hat{b}_M(d,j))_k=0} \sqrt{n}(c^M(d,j,\gamma))_k - (A^M(d,j,\gamma)z)_k$$

for $M=L,U$.

Under Assumptions 3.3 and 3.4, the distribution of $\sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p})$ can be approximated by a $\mathcal{N}(\sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}p), \ell_{d,j_L^*}\Sigma\ell'_{d,j_L^*})$ distribution asymptotically and $\widehat{Z}_L(d,j_L^*)$ is asymptotically independent. For $\alpha \in (0,1)$, define $\widehat{L}(d)_\alpha^C$ to solve

$$F_{TN}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}\hat{p}); \mu, \ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell'_{d,\hat{j}_L(d)}, \widehat{\mathcal{V}}_L^-\Big(\widehat{Z}_L(d,\hat{j}_L(d)),d,\hat{j}_L(d),\hat{\gamma}_L(d)\Big),$$
$$\widehat{\mathcal{V}}_L^+\Big(\widehat{Z}_L(d,\hat{j}_L(d)),d,\hat{j}_L(d),\hat{\gamma}_L(d)\Big)\Big) = 1-\alpha$$

in μ. Similarly, define $\widehat{U}(d)_\alpha^C$ to solve

$$F_{TN}\Big(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)} + u_{d,\hat{j}_U(d)}\hat{p}); \mu, u_{d,\hat{j}_U(d)}\widehat{\Sigma}u'_{d,\hat{j}_U(d)}, \widehat{\mathcal{V}}_U^-\Big(\widehat{Z}_U(d,\hat{j}_U(d)),d,\hat{j}_U(d),\hat{\gamma}_U(d)\Big),$$
$$\widehat{\mathcal{V}}_U^+\Big(\widehat{Z}_U(d,\hat{j}_U(d)),d,\hat{j}_U(d),\hat{\gamma}_U(d)\Big)\Big) = 1-\alpha$$

in μ. Results in Pfanzagl (1994) imply that $\widehat{L}(d)_\alpha^C$ and $\widehat{U}(d)_\alpha^C$ are optimal α quantile-unbiased estimators of $\sqrt{n}(\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}p)$ and $\sqrt{n}(\tilde{u}_{d,j_L^*} + u_{d,j_L^*}p)$ asymptotically.

Finally, combine these quantile-unbiased estimators to form a conditional CI for the identified interval $[L(d),U(d)]$:

$$(n^{-1/2}\widehat{L}(d)_{\alpha_1}^C, n^{-1/2}\widehat{U}(d)_{1-\alpha_2}^C). \tag{5.1}$$

We establish the conditional uniform asymptotic validity of this CI.

Theorem 5.1. *Suppose Assumptions 3.1–3.5 hold. Then, for any $d \in \{0,...,T\}$ and $0 < \alpha_1, \alpha_2 < 1/2$,*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left\{ \left[\mathbb{P}\Big([L(d), U(d)] \subseteq \Big(n^{-1/2}\widehat{L}(d)^C_{\alpha_1}, n^{-1/2}\widehat{U}(d)^C_{1-\alpha_2} \Big) \Big| d \in \widehat{\mathcal{D}} \Big) - (1-\alpha_1-\alpha_2) \right] \right.$$
$$\left. \cdot \mathbb{P}(d \in \widehat{\mathcal{D}}) \right\} \geq 0$$

for all $d \in \{0,...,T\}$.

6 Unconditional Confidence Intervals

In parallel with the previous section, we now generalize the unconditional CI constructions described in Section 2.3 to the general framework of Section 3. Here, we would like to construct CIs that unconditionally cover the identified interval corresponding to a unique data-dependent object of inferential interest. As mentioned in Section 2.3 if only unconditional coverage of $[L(d), U(d)]$ is desired, the conditional confidence interval (5.1) can be unnecessarily wide. We describe two different methods to form the unconditional probabilistic bounds that constitute the endpoints of these unconditional CIs in this general framework.

The general formation of the probabilistic bounds based upon projecting simultaneous confidence bounds for all possible values of $L(\hat{d})$ and $U(\hat{d})$ proceeds by computing $\hat{c}_{1-\alpha,M}$, the $1-\alpha$ quantile of $\max_{i\in\{1,...,(T+1)J_M\}} \hat{\zeta}_{M,i}/\sqrt{\widehat{\Sigma}_{M,i}}$, where $\hat{\zeta}_M \sim \mathcal{N}(0, \widehat{\Sigma}_M)$ for $M = L, U$ with $\widehat{\Sigma}_L = \ell^{mat}\widehat{\Sigma}\ell^{mat\prime}$, $\widehat{\Sigma}_U = u^{mat}\widehat{\Sigma}u^{mat\prime}$, $\ell^{mat} = (\ell_{0,1}, ..., \ell_{0,J_L}, ..., \ell_{T,1}, ..., \ell_{T,J_L})'$ and $u^{mat} = (u_{0,1}, ..., u_{0,J_U}, ..., u_{T,1}, ..., u_{T,J_U})'$. The lower level $1-\alpha$ projection confidence bound for $\sqrt{n}L(\hat{d})$ is $\widehat{L}(\hat{d})^P_\alpha = \sqrt{n}\widehat{L}(\hat{d}) - \sqrt{\widehat{\Sigma}_{L,\hat{d}J_L+\hat{j}_L(\hat{d})}} \hat{c}_{1-\alpha,L}$ and the upper level $1-\alpha$ projection confidence bound for $\sqrt{n}U(\hat{d})$ is $\widehat{U}^P_{1-\alpha}(\hat{d}) = \sqrt{n}\widehat{U}(\hat{d}) + \sqrt{\widehat{\Sigma}_{U,\hat{d}J_U+\hat{j}_U(\hat{d})}} \hat{c}_{1-\alpha,U}$.

Combining these two confidence bounds at appropriate levels yields an unconditional CI for the identified interval $[L(\hat{d}), U(\hat{d})]$ of the selected $\hat{d}$,

$$(n^{-1/2}\widehat{L}(\hat{d})^P_{\alpha_1}, n^{-1/2}\widehat{U}(\hat{d})^P_{1-\alpha_2}), \tag{6.1}$$

with uniformly correct asymptotic coverage, regardless of how $\hat{d}$ is selected from the data.

Theorem 6.1. *Suppose Assumptions 3.1 and 3.3–3.5 hold. Then, for any (random) $\hat{d} \in \{0,...,T\}$ and $0 < \alpha_1, \alpha_2 < 1/2$,*

$$\liminf_{n \to \infty} \inf_{\mathbb{P} \in \mathcal{P}_n} \mathbb{P}\left([L(\hat{d}), U(\hat{d})] \subseteq \left(n^{-1/2} \widehat{L}(\hat{d})_{\alpha_1}^P, n^{-1/2} \widehat{U}(\hat{d})_{1-\alpha_2}^P \right) \right) \geq 1 - \alpha_1 - \alpha_2.$$

Note that the projection CI (6.1) has the benefit of correct coverage regardless of how $\hat{d}$ is chosen from the data. In this sense, it is more robust than the other CIs we propose in this paper. On the other hand, by using the common selection structure of Assumption 3.2, we are able to produce a hybrid CI that combines the strengths of the conditional CI (5.1) and the projection CI (6.1) which, as described in Section 2.3, are shorter under complementary scenarios.

In analogy with the construction of the conditional CIs, to construct the hybrid CIs we begin by characterizing the conditional distributions of $\widehat{L}(\hat{d})$ and $\widehat{U}(\hat{d})$ but now adding an additional component to the conditioning events. More specifically, under Assumptions 3.2 and 3.6, by intersecting the events

$$
\begin{aligned}
&\left\{ \hat{d} = d^*, \hat{j}_L(\hat{d}) = j_L^*, \hat{\gamma}_L(\hat{d}) = \gamma_L^* \right\} = \left\{ d^* \in \widehat{\mathcal{D}}, \hat{j}_L(d^*) = j_L^*, \hat{\gamma}_L(d^*) = \gamma_L^* \right\} \\
&= \left\{ \widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d^*, j_L^*), d^*, j_L^*, \gamma_L^* \right) \leq \sqrt{n}(\tilde{\ell}_{d^*, j_L^*} + \ell_{d^*, j_L^*} \hat{p}) \leq \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L(d^*, j_L^*), d^*, j_L^*, \gamma_L^* \right), \right. \\
&\qquad\qquad\qquad \left. \widehat{\mathcal{V}}_L^0\left(\widehat{Z}_L(d^*, j_L^*), d^*, j_L^*, \gamma_L^* \right) \geq 0 \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\left\{ \sqrt{n}(\tilde{\ell}_{\hat{d}, \hat{j}_L(\hat{d})} + \ell_{\hat{d}, \hat{j}_L(\hat{d})} p) \geq \widehat{L}_\beta^P(\hat{d}) \right\} \\
&= \left\{ \sqrt{n}(\tilde{\ell}_{\hat{d}, \hat{j}_L(\hat{d})} + \ell_{\hat{d}, \hat{j}_L(\hat{d})} \hat{p}) \leq \sqrt{n}(\tilde{\ell}_{\hat{d}, \hat{j}_L(\hat{d})} + \ell_{\hat{d}, \hat{j}_L(\hat{d})} p) + \hat{c}_{1-\beta, L} \sqrt{\widehat{\Sigma}_{L, \hat{d} J_L + \hat{j}_L(\hat{d})}} \right\}
\end{aligned}
$$

for some $0<\beta<\alpha<1$, we have

$$\sqrt{n}\widehat{L}(\hat{d})\Big|\Big\{\hat{d}=d^*,\hat{j}_L(\hat{d})=j_L^*,\hat{\gamma}_L(\hat{d})=\gamma_L^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}_\beta^P(\hat{d})\Big\}$$

$$\sim\sqrt{n}(\tilde{\ell}_{d^*,j_L^*}+\ell_{d^*,j_L^*}\hat{p})\Big|\Big\{\widehat{\mathcal{V}}_L^-\Big(\widehat{Z}_L(d^*,j_L^*),d^*,j_L^*,\gamma_L^*\Big)\leq\sqrt{n}(\tilde{\ell}_{d^*,j_L^*}+\ell_{d^*,j_L^*}\hat{p})$$

$$\leq\widehat{\mathcal{V}}_L^{+,H}\Big(\widehat{Z}_L(d^*,j_L^*),d^*,j_L^*,\gamma_L^*,\sqrt{n}(\tilde{\ell}_{d^*,j_L^*}+\ell_{d^*,j_L^*}p)\Big),\widehat{\mathcal{V}}_L^0\Big(\widehat{Z}_L(d^*,j_L^*),d^*,j_L^*,\gamma_L^*\Big)\geq0\Big\},$$

where

$$\widehat{\mathcal{V}}_L^{+,H}(z,d,j,\gamma,\mu)=\min\Big\{\widehat{\mathcal{V}}_L^+(z,d,j,\gamma),\mu+\hat{c}_{1-\beta,L}\sqrt{\widehat{\Sigma}_{L,dJ_L+j}}\Big\}.$$

Similarly,

$$\sqrt{n}\widehat{U}(\hat{d})\Big|\Big\{\hat{d}=d^*,\hat{j}_U(\hat{d})=j_U^*,\hat{\gamma}_U(\hat{d})=\gamma_U^*,\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)\leq\widehat{U}_{1-\beta}^P(\hat{d})\Big\}$$

$$\sim\sqrt{n}(\tilde{u}_{d^*,j_U^*}+u_{d^*,j_U^*}\hat{p})\Big|\Big\{\widehat{\mathcal{V}}_U^{-,H}\Big(\widehat{Z}_U(d^*,j_U^*),d^*,j_U^*,\gamma_U^*,\sqrt{n}(\tilde{u}_{d^*,j_U^*}+u_{d^*,j_U^*}p)\Big)$$

$$\leq\sqrt{n}(\tilde{u}_{d^*,j_U^*}+u_{d^*,j_U^*}\hat{p})\leq\widehat{\mathcal{V}}_U^+\Big(\widehat{Z}_U(d^*,j_U^*),d^*,j_U^*,\gamma_U^*\Big),\widehat{\mathcal{V}}_U^0\Big(\widehat{Z}_U(d^*,j_U^*),d^*,j_U^*,\gamma_U^*\Big)\geq0\Big\},$$

where

$$\widehat{\mathcal{V}}_U^{-,H}(z,d,j,\gamma,\mu)=\max\Big\{\widehat{\mathcal{V}}_U^-(z,d,j,\gamma),\mu-\hat{c}_{1-\beta,U}\sqrt{\widehat{\Sigma}_{U,dJ_U+j}}\Big\}.$$

Since the distribution of $\sqrt{n}(\tilde{\ell}_{d^*,j_L^*}+\ell_{d^*,j_L^*}\hat{p})$ can be approximated by a $\mathcal{N}(\sqrt{n}(\tilde{\ell}_{d^*,j_L^*}+\ell_{d^*,j_L^*}p),\ell_{d^*,j_L^*}\Sigma\ell_{d^*,j_L^*}')$ distribution asymptotically and $\widehat{Z}_L(d^*,j_L^*)$ is asymptotically independent, we again work with the truncated normal distribution to compute a hybrid probabilistic lower bound for $\sqrt{n}L(\hat{d})$: for $0<\beta<\alpha<1$, define $\widehat{L}(d)_\alpha^H$ to solve

$$F_{TN}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}\hat{p});\mu,\ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell_{d,\hat{j}_L(d)}',\widehat{\mathcal{V}}_L^-\Big(\widehat{Z}_L(d,\hat{j}_L(d)),d,\hat{j}_L(d),\hat{\gamma}_L(d)\Big),$$

$$\widehat{\mathcal{V}}_L^{+,H}\Big(\widehat{Z}_L(d,\hat{j}_L(d)),d,\hat{j}_L(d),\hat{\gamma}_L(d),\mu\Big)\Big)=\frac{1-\alpha}{1-\beta}.$$

in μ. Similarly, define $\widehat{U}(d)_\alpha^H$ to solve

$$F_{TN}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}\hat{p});\mu,u_{d,\hat{j}_U(d)}\widehat{\Sigma}u'_{d,\hat{j}_U(d)},\widehat{\mathcal{V}}_U^{-,H}\left(\widehat{Z}_U(d,\hat{j}_U(d)),d,\hat{j}_U(d),\hat{\gamma}_U(d),\mu\right),\right.$$
$$\left.\widehat{\mathcal{V}}_U^{+}\left(\widehat{Z}_U(d,\hat{j}_U(d)),d,\hat{j}_U(d),\hat{\gamma}_U(d)\right)\right)=\frac{1-\alpha}{1-\beta}$$

in μ.

Here, $\widehat{L}(\hat{d})_\alpha^H$ is an unconditionally valid probabilistic lower bound for $L(\hat{d})$ and $\widehat{U}(\hat{d})_{1-\alpha}^H$ is an unconditionally valid probabilistic upper bound for $U(\hat{d})$. Combining these two confidence bounds at appropriate levels yields an unconditional CI for the identified interval $[L(\hat{d}),U(\hat{d})]$ of the selected $\hat{d}$,

$$(n^{-1/2}\widehat{L}(\hat{d})_{\alpha_1}^H,n^{-1/2}\widehat{U}(\hat{d})_{1-\alpha_2}^H), \tag{6.2}$$

with uniformly correct asymptotic coverage when $\hat{d}$ is selected from the data according to Assumptions 3.2 and 3.6.

Theorem 6.2. *Suppose Assumptions 3.1–3.6 hold. Then, for any $0<\alpha_1,\alpha_2<1/2$,*

$$\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left([L(\hat{d}),U(\hat{d})]\subseteq\left(n^{-1/2}\widehat{L}(\hat{d})_{\alpha_1}^H,n^{-1/2}\widehat{U}(\hat{d})_{1-\alpha_2}^H\right)\right)\geq 1-\alpha_1-\alpha_2.$$

7 Reality Check Power Comparison

To our knowledge, the CIs proposed in this paper are the first with proven asymptotic validity for interval-identified parameters selected from the data. Therefore, we have no existing inference method to compare the performance of our CIs to when the interval-identified parameter is data-dependent. However, as discussed in Section 3 above, our inference framework covers cases for which the interval-identified parameter is chosen a priori. For these special cases, there is a large literature on inference on partially-identified parameters or their identified sets that can be applied to form CIs. Although these special cases are not of primary interest for this paper, in this section we compare the performance of our proposed inference methods to one of the leading inference methods in the partial identification literature as a "reality check" on whether

our proposed methods are reasonably informative.

In particular, we compare the power of a test implied by our hybrid CI (i.e., a test that rejects when the value of the parameter under the null hypothesis lies outside of the hybrid CI) to the power of the hybrid test of Andrews et al. (2023), which applies to a general class of moment-inequality models. When d is chosen a priori and the parameter p is equal to a vector of moments of underlying data, Assumption 3.1 implies that $L(d) \leq W(d) \leq U(d)$ can be written as a set of (unconditional) moment inequalities, a special case of the general framework of that paper. Of the many papers on inference for moment inequalities, we choose the test of Andrews et al. (2023) for comparison for two reasons: (i) it has been shown to be quite competitive in terms of power and (ii) it is also based upon a (different) inference method that is a hybrid between conditional and projection-based inference.

We compare the power of tests on the ATE of the treatment in the same setting of the Manski bounds example of Section 2, strengthening the mean independence assumption $E(Y(d)|Z) = E(Y(d))$ to full statistical independence $(Y(1), Y(0)) \perp Z$ and using the sharp bounds on the ATE $W(1) = E(Y(1)) - E(Y(0))$ derived by Balke and Pearl (1997, 2011):

$$L(1) = \max \begin{cases} p^{111} + p^{000} - 1 \\ p^{110} + p^{001} - 1 \\ p^{110} - p^{111} - p^{101} - p^{010} - p^{100} \\ p^{111} - p^{110} - p^{100} - p^{011} - p^{101} \\ -p^{011} - p^{101} \\ -p^{010} - p^{100} \\ p^{001} - p^{011} - p^{101} - p^{010} - p^{000} \\ p^{000} - p^{010} - p^{100} - p^{011} - p^{001} \end{cases}$$
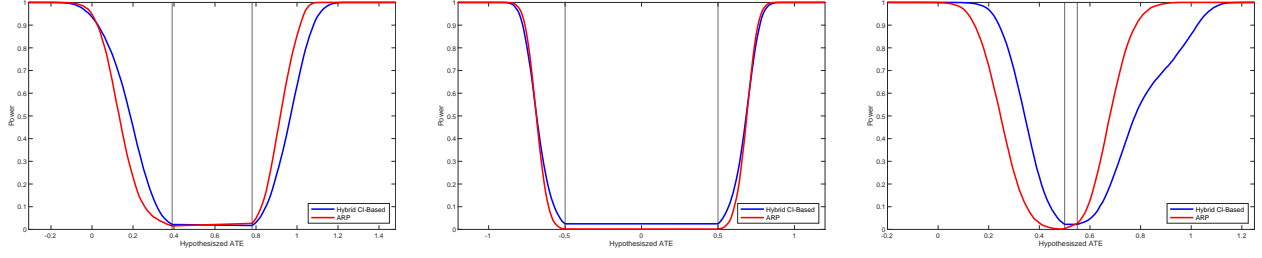
Figure 1: Power curves for hybrid Andrews et al. (2023) test (red) and test implied by hybrid CI (blue) of ATE using bounds from Balke and Pearl (1997, 2011) for $p=(.08,.001,.001,.073,.139,.473)'$ (left), $p=(.25,.25,.25,.25,.25,.25)'$ (middle) and $p=(.01,.44,.01,.01,.01,.54)'$ (right) and $n=100$. The vertical lines illustrate the true identified interval in each of these settings.

and

$$
U(1) = \min \left\{
\begin{array}{c}
1 - p^{011} - p^{100} \\
1 - p^{010} - p^{101} \\
-p^{010} + p^{011} + p^{001} + p^{110} + p^{000} \\
-p^{011} + p^{111} + p^{001} + p^{010} + p^{000} \\
p^{111} + p^{001} \\
p^{110} + p^{000} \\
-p^{101} + p^{111} + p^{001} + p^{110} + p^{100} \\
-p^{100} + p^{110} + p^{000} + p^{111} + p^{101}
\end{array}
\right\}.
$$

For a sample size of $n=100$, Figure 1 plots the power curves of the hybrid Andrews et al. (2023) test and the test implied by our hybrid CI for three different DGPs within the framework of this example, as well as the true identified interval for the ATE. The DGP corresponding to $p=(.08,.001,.001,.073,.139,.473)'$ is calibrated to the probabilities estimated by Balke and Pearl (2011) in the context of a treatment for high-cholesterol (specifically, by the drug cholestyramine).[5] The DGP corresponding to $p=(.25,.25,.25,.25,.25,.25)'$ generates completely uninformative bounds for the ATE. And the DGP corresponding to $p=(.01,.44,.01,.01,.01,.54)'$ generates quite informative bounds.

We can see that overall, the power of the test implied by our hybrid CI is quite competitive with that of Andrews et al. (2023). Interestingly, it appears that our test tends to be more powerful than

---

[5]Balke and Pearl (2011) estimate p to equal $(.081,0,0,.073,.139,.473)'$. If the true DGP is set exactly equal to this, Assumption 3.5 would be violated. We therefore slightly alter the calibrated probabilities.

that of Andrews et al. (2023) when the true ATE is larger than the hypothesized one, which can be seen from the hypothesized ATE lying to the left of the lower bound of the identified interval, and vice versa. Although the main innovation of our inference procedures is really their validity in the presence of data-dependent selection, the exercise of this section is reassuring for the informativeness of the procedures we propose as they are quite competitive in the absence of selection.

8   Finite Sample Performance of Confidence Intervals

Moving now to a context in which the object of interest is selected from the data, we compare the finite sample performance of our conditional, projection and hybrid CIs again in the setting of the Manski bounds example of Section 2. In this case, we are interested in inference on the average potential outcome $W(\hat{d})$ for $W(d) = \mathbb{E}[Y(d)]$, where interest arises either in the average potential outcome for treatment $(\hat{d}=1)$ or control $(\hat{d}=0)$ depending upon which has the largest estimated lower bound: $\hat{d} = \text{argmax}_{d \in \{0,1\}} \widehat{L}(d)$. This form of $\hat{d}$ corresponds to case 1. of Proposition 3.1 and we use the corresponding result of the proposition to specify $\hat{\gamma}_L(\hat{d})$ and $\hat{\gamma}_U(d)$ in the construction of the conditional and hybrid CIs.

For the same DGPs as in Section 7, we compute the unconditional coverage frequencies of the conditional, projection and hybrid CIs as well as that of the conventional CI based upon the normal distribution. These coverage frequencies are reported in Table 1. Consistent with the asymptotic results of Theorems 5.1–6.2, the conditional, projection and hybrid CIs all have correct coverage for all DGPs and the modest sample size of $n=100$. Also consistent with Theorem 6.1, we note that the projection CI tends to be conservative with true coverage above the nominal level of 95%. Finally, we note that the conventional CI can substantially under-cover, consistent with the discussion in Section 2.1.

Next, we compare the length quantiles of the CIs with correct coverage for these same DGPs. Figure 2 plots the ratios of the 5th, 25th, 50th, 75th and 95th quantiles of the length of the conditional, projection and hybrid CIs relative to those same length quantiles of the projection CI. As can be seen from the figure, the conditional CI has the tendency to become very long, especially at high quantile levels, for certain DGPs, whereas the hybrid CI tends to perform the best overall

36

Table 1: Unconditional Coverage Frequencies

Data-Generating Process	Confidence Interval			
	Conv	Cond	Proj	Hyb
$p = (.08, .001, .001, .073, .139, .473)'$	0.95	0.95	0.99	0.95
$p = (.25, .25, .25, .25, .25, .25)'$	0.85	0.95	0.96	0.95
$p = (.01, .44, .01, .01, .01, .54)'$	0.95	0.95	0.99	0.95

This table reports unconditional coverage frequencies for the potential outcome selected by maximizing the estimated lower bound on the potential outcomes of either treatment or control, all evaluated at the nominal coverage level of 95%. Coverage frequencies are reported for conventional ("Conv"), conditional ("Cond"), projection ("Proj") and hybrid ("Hyb") confidence intervals for a sample size of $n = 100$.
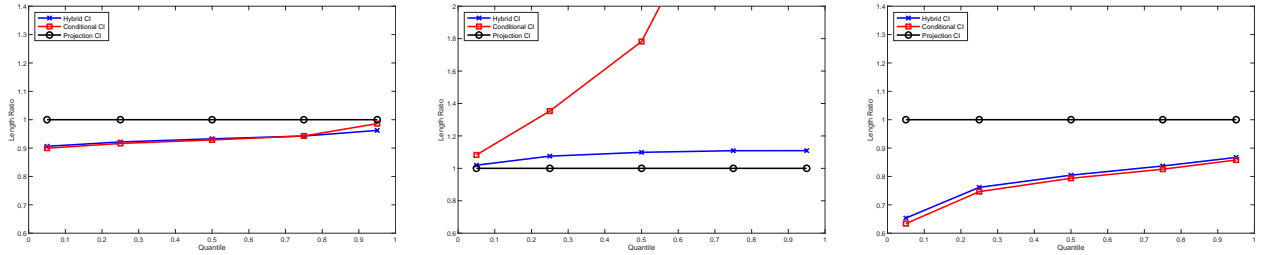
Figure 2: Ratios of CI length quantiles relative to those of the projection CI for the conditional CI (red), projection CI (black) and hybrid CI (blue) of the average potential outcome selected by maximizing the estimated lower bound when $p = (.08, .001, .001, .073, .139, .473)'$ (left), $p = (.25, .25, .25, .25, .25, .25)'$ (middle) and $p = (.01, .44, .01, .01, .01, .54)'$ (right) and $n = 100$.

by limiting the worst-case length performance of the conditional CI relative to the projection CI. Relative to projection, the hybrid CI enjoys length reductions of 20-30% for favorable DGPs while only showing length increases of 5-10% for unfavorable DGPs.

9   Technical Appendix

Proof of Proposition 3.1: Assumption 3.6 is trivially satisfied by supposition. For Assumption 3.2, note that $d \in \widehat{\mathcal{D}}$ is equivalent to

$$w_L \widehat{L}(d) + w_U \widehat{U}(d) \geq w_L \widehat{L}(d') + w_U \widehat{U}(d')$$

for all $d' \in \{0,...,T\}$ or

$$w_L \cdot \max_{j \in \{1,...,J_L\}} \{\tilde{\ell}_{d,j} + \ell_{d,j}\hat{p}\} + w_u \cdot \min_{j \in \{1,...,J_U\}} \{\tilde{u}_{d,j} + u_{d,j}\hat{p}\}$$

$$\geq w_L \cdot \max_{j \in \{1,...,J_L\}} \{\tilde{\ell}_{d',j} + \ell_{d',j}\hat{p}\} + w_u \cdot \min_{j \in \{1,...,J_U\}} \{\tilde{u}_{d',j} + u_{d',j}\hat{p}\}$$

for all $d' \in \{0,...,T\}$. Therefore, we have the following.

1. When $w_U = 0$, $d \in \widehat{\mathcal{D}}$ and $\hat{j}_L(d) = j_L^*$ if and only if $\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p} \geq \tilde{\ell}_{d',j} + \ell_{d',j}\hat{p}$ for all $d' \in \{0,...,T\}$ and $j \in \{1,...,J_L\}$. Similarly, when $w_U = 0$, $d \in \widehat{\mathcal{D}}$, $\hat{j}_U(d) = j_U^*$ and $\hat{j}_L(d) = j_L^*$ if and only if $\tilde{\ell}_{d,j_L^*} + \ell_{d,j_L^*}\hat{p} \geq \tilde{\ell}_{d',j} + \ell_{d',j}\hat{p}$ for all $d' \in \{0,...,T\}$ and $j \in \{1,...,J_L\}$ and $A_U(d,j_U^*)\hat{p} \leq c_U(d,j_U^*)$.

2. When $w_L = 0$, $d \in \widehat{\mathcal{D}}$, $(\hat{j}_U(0),...,\hat{j}_U(T))' = (j_U^*(0),...,j_U^*(T))'$ and $\hat{j}_L(d) = j_L^*$ if and only if $\tilde{u}_{d,j_U^*(d)} + u_{d,j_U^*(d)}\hat{p} \geq \tilde{u}_{d',j_U^*(d')} + u_{d',j_U^*(d')}\hat{p}$ for all $d' \in \{0,...,T\}$, $A_U(d',j_U^*(d'))\hat{p} \leq c_U(d',j_U^*(d'))$ for all $d' \in \{0,...,T\}$ and $A_L(d,j_L^*)\hat{p} \leq c_L(d,j_L^*)$. Similarly, when $w_L = 0$, $d \in \widehat{\mathcal{D}}$ and $(\hat{j}_U(0),...,\hat{j}_U(T))' = (j_U^*(0),...,j_U^*(T))'$ if and only if $\tilde{u}_{d,j_U^*(d)} + u_{d,j_U^*(d)}\hat{p} \geq \tilde{u}_{d',j_U^*(d')} + u_{d',j_U^*(d')}\hat{p}$ and $A_U(d',j_U^*(d'))\hat{p} \leq c_U(d',j_U^*(d'))$ for all $d' \in \{0,...,T\}$.

3. When $w_L, w_U \neq 0$, $d \in \widehat{\mathcal{D}}$ and

$$(\hat{j}_L(0),...,\hat{j}_L(T),\hat{j}_U(0),...,\hat{j}_U(T))' = (j_L^*(0),...,j_L^*(T),j_U^*(0),...,j_U^*(T))'$$

if and only if

$$w_L(\tilde{\ell}_{d,j_L^*(d)} + \ell_{d,j_L^*(d)}\hat{p}) + w_U(\tilde{u}_{d,j_U^*(d)} + u_{d,j_U^*(d)}\hat{p})$$

$$\geq w_L(\tilde{\ell}_{d',j_L^*(d')} + \ell_{d',j_L^*(d')}\hat{p}) + w_U(\tilde{u}_{d',j_U^*(d')} + u_{d',j_U^*(d')}\hat{p}),$$

$A_U(d',j_U^*(d'))\hat{p} \leq c_U(d',j_U^*(d'))$ and $A_L(d',j_L^*(d'))\hat{p} \leq c_L(d',j_L^*(d'))$ for all $d' \in \{0,...,T\}$. ∎

Proof of Proposition 4.1: Assumption 3.6 is trivially satisfied by supposition. For Assumption 3.2, note that

38

1. $\underline{k}_i = \underline{k}_i^*$ and $\bar{k}_i = \bar{k}_i^*$ if and only if $(\mathfrak{u}_{\underline{k}_i^*} - \mathfrak{u}_k)\hat{p} \leq \tilde{\mathfrak{u}}_k - \tilde{\mathfrak{u}}_{\underline{k}_i^*} + (\mathfrak{u}_k - \mathfrak{u}_{\underline{k}_i^*})\varepsilon_i$ for all $k = 1,\dots,K_u$ and $(\mathfrak{l}_k - \mathfrak{l}_{\bar{k}_i^*})\hat{p} \leq \tilde{\mathfrak{l}}_{\bar{k}_i^*} - \tilde{\mathfrak{l}}_k + (\mathfrak{l}_{\bar{k}_i^*} - \mathfrak{l}_k)\varepsilon_i$ for all $k = 1,\dots,K_L$;

2. $s_i^{\mathfrak{l}} = -$ and $s_i^{\mathfrak{u}} = -$ if and only if $\mathfrak{l}_{\bar{k}_i}\hat{p} \leq -\tilde{\mathfrak{l}}_{\bar{k}_i} - \mathfrak{l}_{\bar{k}_i}\varepsilon_i$ and $\mathfrak{u}_{\underline{k}_i}\hat{p} \leq -\tilde{\mathfrak{u}}_{\underline{k}_i} - \mathfrak{u}_{\underline{k}_i}\varepsilon_i$;

3. $\hat{d} = 1$ if and only if $\sum_{i\in\underline{m}}(-\mathfrak{u}_{\underline{k}_i}\hat{p}) + \sum_{i\in\bar{m}}(-\mathfrak{l}_{\bar{k}_i}\hat{p}) \leq \sum_{i\in\underline{m}}(\tilde{\mathfrak{u}}_{\underline{k}_i} + \mathfrak{u}_{\underline{k}_i}\varepsilon_i) + \sum_{i\in\bar{m}}(\tilde{\mathfrak{l}}_{\bar{k}_i} + \mathfrak{l}_{\bar{k}_i}\varepsilon_i)$, where $\underline{m} = \{i\in\{1,\dots,m\} : s_i^{\mathfrak{u}} = +\}$ and $\bar{m} = \{i\in\{1,\dots,m\} : s_i^{\mathfrak{l}} = +\}$. $\blacksquare$

Lemma 9.1. *Suppose Assumptions 3.1–3.5 hold. Then, for any $0 < \alpha < 1$,*

$$\limsup_{n\to\infty}{}_{\mathbb{P}\in\mathcal{P}_n}\left| \mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C \,\middle|\, d\in\widehat{\mathcal{D}}, \hat{j}_L(d) = j_L^*, \hat{\gamma}_L(d) = \gamma_L^* \right) - \alpha \right|$$

$$\cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}}, \hat{j}_L(d) = j_L^*, \hat{\gamma}_L(d) = \gamma_L^* \right) = 0, \tag{9.1}$$

$$\limsup_{n\to\infty}{}_{\mathbb{P}\in\mathcal{P}_n}\left| \mathbb{P}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)} + u_{d,\hat{j}_U(d)}p) \leq \widehat{U}(d)_\alpha^C \,\middle|\, d\in\widehat{\mathcal{D}}, \hat{j}_U(d) = j_U^*, \hat{\gamma}_U(d) = \gamma_U^* \right) - \alpha \right|$$

$$\cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}}, \hat{j}_U(d) = j_U^*, \hat{\gamma}_U(d) = \gamma_U^* \right) = 0, \tag{9.2}$$

for all $d\in\{0,\dots,T\}$, $j_L^*\in\{1,\dots,J_L\}$, $j_U^*\in\{1,\dots,J_U\}$, γ_L^* in the support of $\hat{\gamma}_L(d)$ and γ_U^* in the support of $\hat{\gamma}_U(d)$.

Proof: The proof of (9.2) is nearly identical to the proof of (9.1) so that we only show the latter. Lemma A.1 of Lee et al. (2016) implies that $F_{TN}(t;\mu,\sigma^2,v^-,v^+)$ is strictly decreasing in μ so that $\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C$ is equivalent to

$$F_{TN}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}\hat{p}); \sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}p), \ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell'_{d,\hat{j}_L(d)}, \widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right), \right.$$

$$\left. \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right) \right) \geq 1 - \alpha,$$

where we use $\widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right)$ and $\widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right)$ as shorthand for

$$\widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d,\hat{j}_L(d)), d, \hat{j}_L(d), \hat{\gamma}_L(d) \right)$$

and

$$\widehat{\mathcal{V}}_L^+ \left(\widehat{Z}_L(d,\hat{j}_L(d)),d,\hat{j}_L(d),\hat{\gamma}_L(d) \right).$$

In addition, Lemma 2 of McCloskey (2024) implies

$$F_{TN}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}\hat{p}); \sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p), \ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell'_{d,\hat{j}_L(d)}, \widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right), \right.$$
$$\left. \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L(d,\hat{j}_L(d)) \right) \right)$$
$$= F_{TN}\left(\ell_{d,\hat{j}_L(d)}\sqrt{n}(\hat{p}-p);0,\ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell'_{d,\hat{j}_L(d)}, \widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L^*(d,\hat{j}_L(d)) \right), \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L^*(d,\hat{j}_L(d)) \right) \right),$$

where $\widehat{Z}_L^*(d,j) = \sqrt{n}\hat{p} - \hat{b}_L(d,j)\ell_{d,j}\sqrt{n}(\hat{p}-p)$. Therefore, $\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C$ is equivalent to

$$F_{TN}\left(\ell_{d,\hat{j}_L(d)}\sqrt{n}(\hat{p}-p);0,\ell_{d,\hat{j}_L(d)}\widehat{\Sigma}\ell'_{d,\hat{j}_L(d)}, \widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L^*(d,\hat{j}_L(d)) \right), \widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L^*(d,\hat{j}_L(d)) \right) \right) \geq 1-\alpha. \qquad (9.3)$$

Under Assumptions 3.1, 3.3 and 3.5, a slight modification of Lemma 5 of Andrews et al. (2024) implies that to prove (9.1), it suffices to show that for all subsequences $\{n_s\} \subset \{n\}$, $\{\mathbb{P}_{n_s}\} \in \times_{n=1}^\infty \mathcal{P}_n$ with

1. $\Sigma(\mathbb{P}_{n_s}) \to \Sigma^* \in \mathcal{S} = \{\Sigma : 1/\bar{\lambda} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \bar{\lambda}\}$,

2. $\mathbb{P}_{n_s}\left(d \in \widehat{\mathcal{D}}, \hat{j}_L(d)=j_L^*, \hat{\gamma}_L(d)=\gamma_L^* \right) \to q^* \in (0,1]$, and

3. $\sqrt{n_s}p_{n_s}(\mathbb{P}_{n_s}) \to p^* \in [0,\infty]^{\dim(\xi)}$

for some finite $\bar{\lambda}$, we have

$$\lim_{n\to\infty} \mathbb{P}_{n_s}\left(\sqrt{n_s}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p(\mathbb{P}_{n_s})) \leq \widehat{L}(d)_\alpha^C \,\Big|\, d \in \widehat{\mathcal{D}}, \hat{j}_L(d)=j_L^*, \hat{\gamma}_L(d)=\gamma_L^* \right) = \alpha$$

for all $d \in \{0,...,T\}$, $j_L^* \in \{1,...,J_L\}$ and γ_L^* in the support of $\hat{\gamma}_L(d)$. Let $\{\mathbb{P}_{n_s}\}$ be a sequence satisfying conditions 1.–3. Under $\{\mathbb{P}_{n_s}\}$, $(\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s})),\widehat{\Sigma}) \xrightarrow{d} (\xi^*,\Sigma^*)$ by Assumptions 3.3 and 3.4, where $\xi^* \sim \mathcal{N}(0,\Sigma^*)$. Note that condition 2. implies $\sqrt{n_s}(c^L(d,j_L^*,\gamma_L^*) - A^L(d,j_L^*,\gamma_L^*)\hat{p})$ is asymptotically greater than zero with positive probability for all $d \in \{0,...,T\}$, $j_L^* \in \{1,...,J_L\}$ and γ_L^* in the support

40

of $\hat{\gamma}_L(d)$ under $\{\mathbb{P}_{n_s}\}$ since $\{d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^*\}=\{c^L(d,j_L^*,\gamma_L^*)-A^L(d,j_L^*,\gamma_L^*)\hat{p}\geq 0\}$ by Assumptions 3.1 and 3.2. Consequently, Assumption 3.3 and condition 3. imply $\sqrt{n_s}(c^L(d,j_L^*,\gamma_L^*)-A^L(d,j_L^*,\gamma_L^*)p(\mathbb{P}_{n_s}))\to\omega(d,j_L^*,\gamma_L^*)>-\infty$ for all $d\in\{0,...,T\}$, $j_L^*\in\{1,...,J_L\}$ and γ_L^* in the support of $\hat{\gamma}_L(d)$. Thus, under Assumptions 3.1–3.5, similar arguments to those used in the proof of Lemma 8 in Andrews et al. (2024) show that for any $d\in\{0,...,T\}$, $j_L^*\in\{1,...,J_L\}$ and γ_L^* in the support of $\hat{\gamma}_L(d)$, $(\widehat{\mathcal{V}}_L^-(\widehat{Z}_L^*(d,j_L^*)),\widehat{\mathcal{V}}_L^+(\widehat{Z}_L^*(d,j_L^*)))\xrightarrow{d}(\mathcal{V}_L^{-,*}(d,j_L^*,\gamma_L^*),\mathcal{V}_L^{+,*}(d,j_L^*,\gamma_L^*))$ under $\{\mathbb{P}_{n_s}\}$, where

$$\mathcal{V}_L^{-,*}(d,j_L^*,\gamma_L^*)=\max_{k:(A^L(d,j_L^*,\gamma_L^*)b_L(d,j_L^*))_k<0}\frac{(\omega(d,j_L^*,\gamma_L^*))_k-(A^L(d,j_L^*,\gamma_L^*)(I-b_L(d,j_L^*)\ell_{d,j_L^*})\xi^*)_k}{(A^L(d,j_L^*,\gamma_L^*)b_L(d,j_L^*))_k},$$
$$\mathcal{V}_L^{+,*}(d,j_L^*,\gamma_L^*)=\min_{k:(A^L(d,j_L^*,\gamma_L^*)b_L(d,j_L^*))_k>0}\frac{(\omega(d,j_L^*,\gamma_L^*))_k-(A^L(d,j_L^*,\gamma_L^*)(I-b_L(d,j_L^*)\ell_{d,j_L^*})\xi^*)_k}{(A^L(d,j_L^*,\gamma_L^*)b_L(d,j_L^*))_k},$$

with $b_L(d,j)=\Sigma^*\ell_{d,j}'\left(\ell_{d,j}\Sigma^*\ell_{d,j}'\right)^{-1}$. This convergence is joint with that of $(\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s})),\widehat{\Sigma})$ so that under $\{\mathbb{P}_{n_s}\}$,

$$\left(\sqrt{n_s}\ell_{d,j_L^*}(\hat{p}-p(\mathbb{P}_{n_s})),\widehat{\Sigma},\widehat{\mathcal{V}}_L^-\left(\widehat{Z}_L^*(d,j_L^*)\right),\widehat{\mathcal{V}}_L^+\left(\widehat{Z}_L^*(d,j_L^*)\right)\right)$$
$$\xrightarrow{d}\left(\ell_{d,j_L^*}\xi^*,\Sigma^*,\mathcal{V}_L^{-,*}(d,j_L^*,\gamma_L^*),\mathcal{V}_L^{+,*}(d,j_L^*,\gamma_L^*)\right) \tag{9.4}$$

for all $d\in\{0,...,T\}$, $j_L^*\in\{1,...,J_L\}$ and γ_L^* in the support of $\hat{\gamma}_L(d)$.

Using (9.4) and the equivalence in (9.3), the remaining arguments to prove (9.1) are nearly identical to those used in the proof of Proposition 1 of McCloskey (2024) and therefore omitted for brevity. ∎

Lemma 9.2. *Suppose Assumptions 3.1–3.5 hold. Then, for any $0<\alpha<1$,*

$$\limsup_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\left|\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p)\leq\widehat{L}(d)_\alpha^C\Big|d\in\widehat{\mathcal{D}}\right)-\alpha\right|\cdot\mathbb{P}\left(d\in\widehat{\mathcal{D}}\right)=0,$$
$$\limsup_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\left|\mathbb{P}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}p)\leq\widehat{U}(d)_\alpha^C\Big|d\in\widehat{\mathcal{D}}\right)-\alpha\right|\cdot\mathbb{P}\left(d\in\widehat{\mathcal{D}}\right)=0,$$

for all $d\in\{0,...,T\}$.

Proof: The results of this lemma follow from Lemma 9.1 since, e.g.,

$$
\limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \left| \mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C \Big| d\in\widehat{\mathcal{D}}\Big) - \alpha \right| \cdot \mathbb{P}\Big(d\in\widehat{\mathcal{D}}\Big)
$$

$$
= \limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \left| \mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C, d\in\widehat{\mathcal{D}}\Big) - \alpha\cdot\mathbb{P}\Big(d\in\widehat{\mathcal{D}}\Big) \right|
$$

$$
= \limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \left| \sum_{j_L^*=1}^{J_L}\sum_{\gamma_L^*} \Big[\mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C, d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) \right.
$$

$$
\left. -\alpha\cdot\mathbb{P}\Big(d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big)\Big] \right|
$$

$$
= \limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \left| \sum_{j_L^*=1}^{J_L}\sum_{\gamma_L^*} \Big[\mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C \Big| d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) - \alpha\Big] \right.
$$

$$
\left. \cdot \mathbb{P}\Big(d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) \right|
$$

$$
\leq \limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \sum_{j_L^*=1}^{J_L}\sum_{\gamma_L^*} \left| \mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C \Big| d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) - \alpha \right|
$$

$$
\cdot \mathbb{P}\Big(d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big)
$$

$$
\leq \sum_{j_L^*=1}^{J_L}\sum_{\gamma_L^*} \limsup_{n\to\infty\,\mathbb{P}\in\mathcal{P}_n} \left| \mathbb{P}\Big(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p) \leq \widehat{L}(d)_\alpha^C \Big| d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) - \alpha \right|
$$

$$
\cdot \mathbb{P}\Big(d\in\widehat{\mathcal{D}},\hat{j}_L(d)=j_L^*,\hat{\gamma}_L(d)=\gamma_L^* \Big) = 0,
$$

where the inner sums $\sum_{\gamma_L^*}$ are over the elements of the support of $\hat{\gamma}_L(d)$. ∎

Proof of Theorem 5.1: The result of this theorem follows from Lemma 9.2 since

$$
\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[\mathbb{P}\left(\sqrt{n}[L(d),U(d)] \subseteq \left(\widehat{L}(d)^C_{\alpha_1}, \widehat{U}(d)^C_{1-\alpha_2} \right) \Big| d\in\widehat{\mathcal{D}} \right) - (1-\alpha_1-\alpha_2) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right)
$$

$$
\geq \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[1 - \mathbb{P}\left(\sqrt{n}L(d)\leq\widehat{L}(d)^C_{\alpha_1} \Big| d\in\widehat{\mathcal{D}} \right) - \mathbb{P}\left(\sqrt{n}U(d)\geq\widehat{U}(d)^C_{1-\alpha_2} \Big| d\in\widehat{\mathcal{D}} \right) \right.
$$

$$
\left. - (1-\alpha_1-\alpha_2) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right)
$$

$$
\geq \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[1 - \mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p)\leq\widehat{L}(d)^C_{\alpha_1} \Big| d\in\widehat{\mathcal{D}} \right) \right.
$$

$$
\left. - \mathbb{P}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}p)\geq\widehat{U}(d)^C_{1-\alpha_2} \Big| d\in\widehat{\mathcal{D}} \right) - (1-\alpha_1-\alpha_2) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right)
$$

$$
= \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[\alpha_1 - \mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p)\leq\widehat{L}(d)^C_{\alpha_1} \Big| d\in\widehat{\mathcal{D}} \right) \right.
$$

$$
\left. + \mathbb{P}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}p)\leq\widehat{U}(d)^C_{1-\alpha_2} \Big| d\in\widehat{\mathcal{D}} \right) - (1-\alpha_2) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right)
$$

$$
\geq \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[\alpha_1 - \mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{d,\hat{j}_L(d)}+\ell_{d,\hat{j}_L(d)}p)\leq\widehat{L}(d)^C_{\alpha_1} \Big| d\in\widehat{\mathcal{D}} \right) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right)
$$

$$
+ \liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \left[\mathbb{P}\left(\sqrt{n}(\tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}p)\leq\widehat{U}(d)^C_{1-\alpha_2} \Big| d\in\widehat{\mathcal{D}} \right) - (1-\alpha_2) \right] \cdot \mathbb{P}\left(d\in\widehat{\mathcal{D}} \right) = 0,
$$

where the second inequality follows from the facts that $L(d) \geq \tilde{\ell}_{d,\hat{j}_L(d)} + \ell_{d,\hat{j}_L(d)}p$ and $U(d) \leq \tilde{u}_{d,\hat{j}_U(d)}+u_{d,\hat{j}_U(d)}p$ almost surely and the final equality follows from Lemma 9.2. ∎

Proof of Theorem 6.1: We start by showing

$$
\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}_n} \mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})^P_{\alpha} \right) \geq 1-\alpha. \tag{9.5}
$$

By the same argument as in the proof of Lemma 9.1, to prove (9.5), it suffices to show

$$
\lim_{n\to\infty} \mathbb{P}_{n_s}\left(\sqrt{n_s}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p(\mathbb{P}_{n_s}))\geq\widehat{L}(\hat{d})^P_{\alpha} \right) \geq 1-\alpha \tag{9.6}
$$

under conditions 1. and 3. Since $\sqrt{n_s}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})^P_{\alpha}$ is equivalent to $\ell_{\hat{d},\hat{j}_L(\hat{d})}\sqrt{n_s}(\hat{p}-p)\leq$

$\hat{c}_{1-\alpha,L}\sqrt{\widehat{\Sigma}_{L,\hat{d}J_L+\hat{j}_L(\hat{d})}}$, the left-hand side of (9.6) is equal to

$$\lim_{n\to\infty}\mathbb{P}_{n_s}\left(\ell_{\hat{d},\hat{j}_L(\hat{d})}\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s}))\leq\hat{c}_{1-\alpha,L}\sqrt{\widehat{\Sigma}_{L,\hat{d}J_L+\hat{j}_L(\hat{d})}}\right)$$

$$\geq\lim_{n\to\infty}\mathbb{P}_{n_s}\left(\ell^{mat}\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s}))\leq\hat{c}_{1-\alpha,L}\sqrt{\mathrm{Diag}\left(\widehat{\Sigma}_L\right)}\right)$$

$$=\mathbb{P}\left(\xi_L\leq c_{1-\alpha,L}\sqrt{\mathrm{Diag}(\Sigma_L)}\right)=\mathbb{P}\left(\max_{i\in\{1,...,(T+1)J_L\}}\frac{\xi_{L,i}}{\sqrt{\Sigma_{L,i}}}\leq c_{1-\alpha,L}\right)=1-\alpha$$

under conditions 1. and 3. for $\xi_L\sim\mathcal{N}(0,\Sigma_L)$, $c_{\alpha,L}$ denoting the α-quantile of

$$\max_{i\in\{1,...,(T+1)J_L\}}\frac{\xi_{L,i}}{\sqrt{\Sigma_{L,i}}}$$

and $\Sigma_L=\ell^{mat}\Sigma\ell^{mat\prime}$, where all inequalities are taken element-wise across vectors, the inequality follows from the fact that $\ell_{\hat{d},\hat{j}_L(\hat{d})}\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s}))$ is a (random) element of $\ell^{mat}\sqrt{n_s}(\hat{p}-p(\mathbb{P}_{n_s}))$ and the first equality follows by identical arguments to those used in the proof of Proposition 11 of Andrews et al. (2024). We have thus proved (9.5). In addition,

$$\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)\leq\widehat{U}(\hat{d})_{1-\alpha}^P\right)\geq1-\alpha.$$

follows by nearly identical arguments. The statement of the theorem then follows by nearly identical arguments to those used in the proof of Theorem 5.1. ∎

Lemma 9.3. *Suppose Assumptions 3.1–3.6 hold. Then, for any $0<\beta<\alpha<1$,*

$$\limsup_{n\to\infty\mathbb{P}\in\mathcal{P}_n}\left|\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\leq\widehat{L}(\hat{d})_\alpha^H\Big|\hat{d}=d^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)-\frac{\alpha-\beta}{1-\beta}\right|$$

$$\cdot\mathbb{P}\left(\hat{d}=d^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)=0, \tag{9.7}$$

$$\limsup_{n\to\infty\mathbb{P}\in\mathcal{P}_n}\left|\mathbb{P}\left(\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)\leq\widehat{U}(\hat{d})_\alpha^H\Big|\hat{d}=d^*,\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)\leq\widehat{U}(\hat{d})_{1-\beta}^P\right)-\frac{\alpha-\beta}{1-\beta}\right|$$

$$\cdot\mathbb{P}(\hat{d}=d^*,\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)\leq\widehat{U}(\hat{d})_{1-\beta}^P)=0, \tag{9.8}$$

for all $d^* \in \{0,...,T\}$.

Proof: The proof of (9.8) is nearly identical to the proof of (9.7) so that we only show the latter. Upon noting that $F_{TN}(t;\mu,\sigma^2,\widehat{\mathcal{V}}_L^-(z,d,j,\gamma),\widehat{\mathcal{V}}_L^{+,H}(z,d,j,\gamma,\mu))$ is decreasing in μ by the same argument used in the proof of Proposition 5 of Andrews et al. (2024) and replacing condition 2. in the proof of Lemma 9.1 with

2'. $\mathbb{P}_{n_s}\left(\hat{d}=d^*,\hat{j}_L(\hat{d})=j_L^*,\hat{\gamma}_L(\hat{d})=\gamma_L^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)\to q^*\in(0,1]$,

completely analogous arguments to those used to prove (9.1) in Lemma 9.1 imply

$$\limsup_{n\to\infty}\sup_{\mathbb{P}\in\mathcal{P}_n}\left|\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\leq\widehat{L}(\hat{d})_\alpha^H\Big|\hat{d}=d^*,\hat{j}_L(\hat{d})=j_L^*,\hat{\gamma}_L(\hat{d})=\gamma_L^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)\right.$$
$$\left.-\frac{\alpha-\beta}{1-\beta}\right|\cdot\mathbb{P}\left(\hat{d}=d^*,\hat{j}_L(\hat{d})=j_L^*,\hat{\gamma}_L(\hat{d})=\gamma_L^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)=0$$

for all $d^* \in \{0,...,T\}$, $j_L^* \in \{1,...,J_L\}$ and γ_L^* in the support of $\hat{\gamma}_L(d^*)$. Then, the same argument as in the proof of Lemma 9.2 implies (9.7). ∎

Lemma 9.4. *Suppose Assumptions 3.1–3.6 hold. Then, for any $0<\alpha<1$,*

$$\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)>\widehat{L}(\hat{d})_\alpha^H\right)\geq 1-\alpha, \tag{9.9}$$

$$\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\sqrt{n}(\tilde{u}_{\hat{d},\hat{j}_U(\hat{d})}+u_{\hat{d},\hat{j}_U(\hat{d})}p)<\widehat{U}(\hat{d})_{1-\alpha}^H\right)\geq 1-\alpha. \tag{9.10}$$

Proof: The proof of (9.10) is nearly identical to the proof of (9.9) so that we only show the latter. Lemma 6 of Andrews et al. (2024) and Lemma 9.3 imply

$$\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)>\widehat{L}(\hat{d})_\alpha^H\right)$$
$$\geq\frac{1-\alpha}{1-\beta}\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\sum_{d^*=0}^{T}\mathbb{P}\left(\hat{d}=d^*,\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)$$
$$=\frac{1-\alpha}{1-\beta}\liminf_{n\to\infty}\inf_{\mathbb{P}\in\mathcal{P}_n}\mathbb{P}\left(\sqrt{n}(\tilde{\ell}_{\hat{d},\hat{j}_L(\hat{d})}+\ell_{\hat{d},\hat{j}_L(\hat{d})}p)\geq\widehat{L}(\hat{d})_\beta^P\right)\geq 1-\alpha,$$

where the final inequality follows from (9.5) in the proof of Theorem 6.1. ∎

Proof of Theorem 6.2: Using Lemma 9.4 in the place of Lemma 9.2, the proof is nearly identical to the proof of Theorem 5.1. ∎

References

Andrews, D. and Guggenberger, P. (2009). Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, 152:19–27. 1

Andrews, I., Kitagawa, T., and McCloskey, A. (2024). Inference on winners. *Quarterly Journal of Economics*, 139:305–358. 1, 1, 2.3, 9, 9, 9, 9, 9

Andrews, I., Roth, J., and Pakes, A. (2023). Inference for linear conditional moment inequalities. *Review of Economic Studies*, 90:2763–2791. 1, 7, 1, 7

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176. 1, 4.2, 7, 1

Balke, A. and Pearl, J. (2011). Nonparametric bounds on causal effects from partial compliance data. UCLA Department of Statistics Paper. 1, 4.2, 7, 1, 5

Christensen, T., Moon, H. R., and Schorfheide, F. (2023). Optimal decision rules when payoffs are partially identified. arXiv:2204.11748v2. 1, 3, 4.6, 4

Cui, Y. and Han, S. (2024). Policy learning with distributional welfare. *arXiv preprint arXiv:2311.15878.* 1, 4.3

D'Adamo, R. (2021). Orthogonal policy learning under ambiguity. *arXiv preprint arXiv:2111.10904.* 1

Han, S. (2024). Optimal dynamic treatment regimes and partial welfare ordering. *Journal of the American Statistical Association (forthcoming).* 1, 4.2, 4.5

Han, S. and Yang, S. (2024). A computational approach to identification of treatment effects for policy evaluation. *Journal of Econometrics (forthcoming).* 1, 4.2

Kallus, N. and Zhou, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890. 1

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86:591–616. 4.3

Kivaranovic, D. and Leeb, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, 116:845–857. 1, 2.3

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. *Annals of Statistics*, 44:907–927. 1, 2.2, 9

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings*, 80:319–323. 1, 2

Manski, C. F. (2021). Econometrics for decision making: building foundations sketched by haavelmo and wald. *Econometrica*, 89:2827–2853. 3

Manski, C. F. (2023). Probabilistic prediction for binary treatment choice: with focus on personalized medicine. *Journal of Econometrics*, 234:647–663. 3

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68:997–1010. 1

McCloskey, A. (2024). Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *Biometrika*, 111:109–127. 1, 2.3, 5, 9, 9

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 80:755–782. 1, 4.2

Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter. 2.2, 5

Pu, H. and Zhang, B. (2021). Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):318–345. 1

Stoye, J. (2012). Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156. 1

Yata, K. (2021). Optimal decision rules under partial identification. *arXiv preprint arXiv:2111.04926*. 1