Partial Identifiability in Discrete Data With Measurement Error Supplementary Material

Noam Finkelstein[*,1] **Roy Adams**[*,2] **Suchi Saria**[1,3] **Ilya Shpitser**[1]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA
[2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, Maryland, USA
[3]Bayesian Health, New York, New York, USA
[*]Equal contributions

1 ESTIMATION AND CONFIDENCE INTERVALS

Due to sampling variation, in real-data applications we know the observed data distribution only up to statistical uncertainty. For any empirical distribution, it is possible to construct an *uncertainty set* for the true observed data distribution. In this section, we present an approach for propagating this uncertainty through the linear program and into the resulting bounds. This approach can be used whenever there is statistical uncertainty about the "right-hand-side" of the constraints of a linear program. As a result, this approach can be used to estimate confidence intervals for the partial identification bounds described in this paper, as well as for bounds on the ATE in the classic IV model.

The two major advantages of this approach over existing methods for calculating uncertainty sets for LP-based estimators [Horowitz and Manski, 2000, Cheng and Small, 2006] are its finite-sample validity, and its computational simplicity. Because the uncertainty is built directly into the linear program, calculating the uncertainty interval does not require resampling or running additional optimization procedures. Perhaps the most common approach in such settings is the bootstrapping method developed in Beran [1990] for balanced simultaneous confidence sets. However, recent work indicates that the bootstrap does not yield valid results for functionals that are not smooth when estimates of the parameters of the distribution are asymptotically Gaussian, as in discrete data [Fang and Santos, 2019]. It is well known that LPs are not smooth in the RHS coefficients of their constraints [Chvatal, 1983].

The basic idea of our approach is that rather than insisting that the full data must marginalize exactly to the empirical distribution, we instead insist only that the full data must marginalize to some distribution in the uncertainty set of the observed data distribution. To facilitate incorporation of such uncertainty sets into the relevant linear programs, we would like to be able to express them linearly. To that end, we make use of the convex polytope uncertainty region developed in Garivier [2011] and succinctly stated in Nowak and Tánczos [2019], described by the expression

$$\mathbb{P}(\mathrm{KL}_{bern}(\hat{p}_i||p_i) \leq \frac{\log(\frac{2k}{\alpha})}{n} \;\; \forall\, i) \geq 1-\alpha, \tag{1}$$

where $\hat{p}_i$ and p_i represent the empirical and true probabilities of outcome i respectively, $\mathrm{KL}_{bern}(p||q)$ is the KL-divergence between Bernoulli distributions with parameters p and q, k represents the number of possible outcomes, and n represents the size of the dataset. In our setting, k is just the product of the cardinalities of the observed variables.

For convenience, we let $\mathcal{U}(\hat{P}, \alpha, n)$ represent the uncertainty polytope at level α around $\hat{P}$ for a sample of size n. The following proposition shows how statistical uncertainty can be incorporated when no assumptions are made about the relationships among the observed proxies $\mathbf{Y}$, as in Section 3.

Proposition 1 (Measurement Error Uncertainty)**.** *Let $\mathcal{M}$ be a convex polytope model for the distribution ϕ of an unobserved random variable X and observed proxies $\mathbf{Y}$, and $\hat{P}_{\mathbf{Y}}$ be the empirical distribution of $\mathbf{Y}$. Define*

$$\mathcal{P} \equiv \{Q : \sum_{x} Q(X=x, \mathbf{Y}) \in \mathcal{U}(\hat{P}_{\mathbf{Y}}, \alpha, n)\}.$$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Then for any functional $\eta(\phi)$,

$$P(\min_{Q \in \mathcal{M} \cap \mathcal{P}} \eta(Q) \leq \eta(\phi) \leq \max_{Q \in \mathcal{M} \cap \mathcal{P}} \eta(Q)) \geq 1 - \alpha.$$

Proof. By the validity of the polytope uncertainty region (1), $P(\phi \in \mathcal{P}) \geq 1 - \alpha$ for any ϕ – including any $\phi \in \mathcal{M}$ – as $\sum_x \phi(X = x, \mathbf{Y}) = P(\mathbf{Y})$. Because $\phi \in \mathcal{M}$ by assumption, $P(\phi \in \mathcal{P}) = P(\phi \in \mathcal{P} \cap \mathcal{M})$. The conclusion then follows trivially. □

Next, we consider uncertainty in IV models, where the observed data constraints are expressed in terms of conditional distributions. For that reason, we create uncertainty polytopes around these conditional distributions, rather than around the full observed data distribution.

Proposition 2 (Uncertainty w/ Instruments)**.** *Let $\mathcal{M}$ be a convex polytope model for ψ, and $\hat{P}_{\mathbf{B}|\mathbf{a}}$ represent the empirical distribution of $P(\mathbf{B} \mid \mathbf{A} = \mathbf{a})$, and $\hat{P}(\mathbf{a})$ represent the empirical probability that $P(\mathbf{A} = \mathbf{a})$. Define*

$$\mathcal{P} \equiv \{Q : \sum_{\tilde{\mathbf{b}}} Q(\tilde{\mathbf{b}}) \prod_{B \in \mathbf{B}} \mathbb{I}\left(\tilde{b}(\mathbf{a}_{Pa(B)}, \mathbf{b}_{Pa(B) \setminus \mathbf{A}}) = \mathbf{b}_B\right) \in \mathcal{U}(\hat{P}_{\mathbf{B}|\mathbf{a}}, 1 - (1-\alpha)^{1/|\mathbf{A}|}, n\hat{P}(\mathbf{a})) \ \forall \, \mathbf{a}\}.$$

Then for any functional $\eta(\psi)$ that is not a function of $P(\mathbf{A})$,

$$P(\min_{Q \in \mathcal{M} \cap \mathcal{P}} \eta(Q) \leq \eta(\psi) \leq \max_{Q \in \mathcal{M} \cap \mathcal{P}} \eta(Q)) \geq 1 - \alpha.$$

Proof. For a fixed $\mathbf{a}$, $P(\mathbf{B} \mid \mathbf{A} = \mathbf{a}) = \sum_{\tilde{\mathbf{b}}} Q(\tilde{\mathbf{b}}) \prod_{B \in \mathbf{B}} \mathbb{I}\left(\tilde{b}(\mathbf{a}_{Pa(B)}, \mathbf{b}_{Pa(B) \setminus \mathbf{A}}) = \mathbf{b}_B\right)$ as described in Section 4. The set $\mathcal{P}$ is therefore the set of distributions over $\tilde{\mathbf{B}}$ such that each observed conditional distribution is in its $1 - (1-\alpha)^{1/|\mathbf{A}|}$ uncertainty polytope, as each such polytope is estimated using a sample of size $n\hat{P}(\mathbf{a})$. Note that estimates of these conditional distributions are statistically independent, as each sample subject is used in the estimation of only the conditional distribution corresponding to the observed value of the instruments in that subject. Therefore the probability that all $|\mathbf{A}|$ of them fall into their $1 - (1-\alpha)^{1/|\mathbf{A}|}$ uncertainty polytopes is simply $1 - \alpha$. The remainder of the proof is as before. □

This approach cannot account for uncertainty when either the objective or the assumptions are a function of $P(\mathbf{A})$. For example, in linear program (6), we target the factual distribution of the unobserved random variable X, which is as the sum of products of the parameters of ψ and $P(\mathbf{A})$. Because the distribution of the instrument is observed, we can in theory substitute in its empirical distribution. However, if we are to account for uncertainty, we must introduce the true parameters of $P(\mathbf{A})$ distribution over the instruments into the program, yielding bi-linear terms. In the following proposition, we provide a graphical criterion for assumptions and objectives that are not functions of $P(\mathbf{A})$, and therefore for which Proposition 2 can be used.

Proposition 3. *The distribution of a set of potential outcomes $Z_1(\mathbf{T_1} = \mathbf{t_1}), \ldots, Z_n(\mathbf{T_N} = \mathbf{t_n})$ is not a function of the distribution of the instruments $P(\mathbf{A})$ if and only if there is no index i such that there is an instrument A with a directed path to Z_i not through $\mathbf{T_i}$.*

2 COMPUTING BOUNDS FOR NON-LINEAR MODELS

Unfortunately, many useful models do not fall into the class of general IV models. For example, the simple Markov chain shown in Figure 1 (f) of the main paper is not a general IV model and, as we will see below, generates non-linear constraints on ϕ. In this section, we describe how non-sharp outer bounds can be derived for such cases. The only known complete procedure for identifying all constraints implied by Bayesian networks on the distribution of a subset of their vertices is an application of quantifier elimination [Gieger and Meek, 1999], which is infeasibly slow for many problems. When constraints are known to exist, for example by Evans' e-separation criterion [Evans, 2012], their exact form may not be known and may not be linear. When constraints are known, but are not linear, it may be possible to derive sharp bounds analytically. For example, the following proposition, proven in Section 3 of the supplementary materials, gives sharp bounds for a three variable Markov chain (Figure 1 (f) of the main paper) over binary variables.

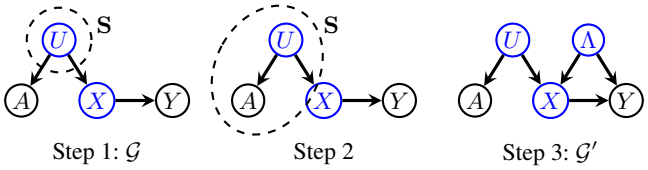

Figure 1: A non-linear graphical model $\mathcal{G}$ and an application of the linear relaxation procedure, resulting in the linear relaxed model $\mathcal{G}'$.

Proposition 4. *Let X and Y be binary variables such that $X \not\perp\!\!\!\perp Y$, let A be a discrete variable such that $A \perp\!\!\!\perp Y|X$ and $P(A = a) > 0$ for all a, and let $p_y = P(Y = y)$ and $p_{y|a} = P(Y = y \mid A = a)$. Then we have the following sharp bounds on $P(X = 1)$:*

$$P(X = 1) \in \bigcup_{y \in \{0,1\}} \left[\frac{p_y - \min_a p_{y|a}}{1 - \min_a p_{y|a}}, \frac{p_y}{\max_a p_{y|a}} \right]. \tag{2}$$

Such analytical bounds, however, are not typically available. In these cases, *non-sharp* bounds can be derived for any graph by relaxing independence assumptions until the model becomes a general IV model. Specifically, any latent variable Bayesian network $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ can be converted to a general IV model $\mathcal{G}'$ by first identifying the set of instruments and then adding a mutual latent confounder between all non-instruments. This procedure is formalized in the following steps:

Procedure 1 (Graphical Relaxation)**.**

1. *Initialize $\mathbf{S}$ as the set of all latent confounders and all variables V such that $Pa(V) = \emptyset$*
2. *For any latent confounder U with exactly two children A and B such that $Pa(A) = \{U\}$ and $Ch(A) = \{B\}$ or $Ch(A) = \emptyset$, add A to $\mathbf{S}$*
3. *Add a latent confounder Λ, and an edge from Λ to each variable in $\mathbf{V}/\mathbf{S}$.*

An example application of this procedure is shown in Figure 1. Because adding an additional latent confound can only *remove* independencies from the graph, this procedure represents relaxations of the constraints on ψ, and therefore enables the calculation of outer bounds for mismeasured variables through linear programming.

Proposition 5. *Suppose $\mathcal{G}'$ is the general IV graph obtained after relaxing a graph $\mathcal{G}$ according to Procedure 1. Then sharp bounds on any linear functional of $P(\mathbf{B_U})$ in $\mathcal{G}'$ will be valid outer bounds in $\mathcal{G}$.*

3 PROOFS

Proof of Proposition 1 of the main paper

First, the constraints places on $P(\mathbf{A}, \mathbf{B})$ by fixing the marginal $P(\mathbf{A}, \mathbf{B_O})$ are trivially linear. Next, as simple consequence of Proposition 5 of Evans [2016], whenever two random variables A and B in a graph $\mathcal{G}$ share a common unobserved parent, and A has no other parents, the model corresponding to the graph is unchanged by removing the common parent and adding an edge from A to B if one does not already exist. As a result, a graph $\mathcal{G}'$ obtained by applying Procedure 1 to any general IV graph $\mathcal{G}$ will represent the same model as $\mathcal{G}$.

In $\mathcal{G}'$, each instrument is randomized with respect to its child in $\mathbf{B}$. Λ can be thought of as a selector such that each variable in $\mathbf{B}$ takes a value determined by Λ and its parents in $\{\mathbf{A}, \mathbf{B}\}$. Consider the *equivalence class* formed by all values of Λ that lead to the same settings of $\mathbf{B}$ for each setting of $\mathbf{A}$ of $[\lambda] = \{\lambda' : P(B \mid \Lambda = \lambda, A = a) = P(B \mid \Lambda = \lambda', A = a) \ \ \forall \ a)$. Then, by [Fine, 1982], when variables are discrete, the conditional distributions $P(\mathbf{B} \mid \mathbf{A})$ can be represented linear combinations of parameters. See Appendix A of Chaves et al. [2017] for a common representation of these linear constraints, in which the linear combination is of probabilities that Λ takes values in various equivalence classes. Note that is equivalent to the response function variable representation we use as a joint setting of the response function variables indicates membership in a particular equivalence class of Λ. Finally, because $P(\mathbf{A}, \mathbf{B_O})$, and by extension $P(\mathbf{A})$, is known, $\mathcal{G}'$ places linear constraints on $P(\mathbf{A}, \mathbf{B})$.

Proof of Proposition 2 of the main paper

First, we recall that any graph $\mathcal{G}$ in general instrumental variable model is equivalent to $\mathcal{G}'$ in which $P(\tilde{\mathbf{B}} = \tilde{\mathbf{b}}, \mathbf{A}) = P(\mathbf{A})\psi_{\tilde{\mathbf{b}}}$. Then,

$$\begin{aligned} P(\mathbf{Z}(\mathbf{t}) = \mathbf{z}) &= \sum_{\mathbf{v}:\mathbf{v_Z}=\mathbf{z}} P(\mathbf{B}(\mathbf{t}) = \mathbf{v_B} \mid \mathbf{A} = \mathbf{v_A}) P(\mathbf{A} = \mathbf{v_A}) \\ &= \sum_{\mathbf{v}:\mathbf{v_Z}=\mathbf{z}} P(\mathbf{B}(\mathbf{t}, \mathbf{v_{V \setminus T}}) = \mathbf{v_B} \mid \mathbf{A} = \mathbf{v_A}) P(\mathbf{A} = \mathbf{v_A}) \\ &= \sum_{\mathbf{v}:\mathbf{v_Z}=\mathbf{z}} P(\mathbf{B}(\mathbf{t}, \mathbf{v_{V \setminus T}}) = \mathbf{v_B}) P(\mathbf{A} = \mathbf{v_A}). \end{aligned}$$

The second equality is by the causal consistency assumption, and the last one is by the independence property of $\mathcal{G}'$. Next, we note that intervention on variables other than the parents of B is irrelevant given intervention on its parents, yielding

$$P(\mathbf{B}(\mathbf{t}, \mathbf{v}) = \mathbf{v_B}) = P(B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = \mathbf{v}_B : B \in \mathbf{B}).$$

Finally, by definition of $\tilde{\mathbf{B}}$, each $B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}})$ is in $\tilde{\mathbf{B}}$, such that $P(B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = v_b : B \in \mathbf{B})$ is simply the probability $\tilde{\mathbf{B}}$ takes a value $\tilde{\mathbf{b}}$ in which $B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = \mathbf{v}_B$ for all $B \in \mathbf{B}$, concluding the proof.

Proof of Proposition 4 of the supplementary materials

Following the approach in Balke and Pearl, we derive bounds on π by translating our assumptions into constraints on π and then find sharp upper (lower) bounds by maximizing (minimizing) π_1 subject to these constraints. We begin with the equality

$$\pi_1 = \frac{p_1 - q_{1|0}}{q_{1|1} - q_{1|0}}. \tag{3}$$

This function is discontinuous at $q_{1|1} = q_{1|0}$ (which is disallowed by assumption), but continuous above and below this line. To derive the bounds in Proposition 4, we take the union of the sharp bounds when $q_{1|1} > q_{1|0}$ and when $q_{1|1} < q_{1|0}$. Consider first the case when $q_{1|1} > q_{1|0}$. For each value a of A we have

$$p_{1|a} = q_{1|0}(1 - \pi_{1|a}) + q_{1|1}\pi_{1|a}$$

Combining this with Equation 3, we can find the sharp upper bound by solving the following (non-linear) optimization problem:

$$\begin{aligned} \max_{q_{1|1} > q_{1|0}} \quad & \frac{p_1 - q_{1|0}}{q_{1|1} - q_{1|0}} \\ \text{s.t.} \quad & p_{1|a} = q_{1|0}(1 - \pi_{1|a}) + q_{1|1}\pi_{1|a} \ \ \forall a \\ & 0 \leq q_{1|0}, q_{1|1}, \pi_{1|a} \leq 1 \ \ \forall a \end{aligned}$$

To solve this optimization problem, we will fix $q_{1|1}$ and optimize with respect to $q_{1|0}$ and then optimize the resulting function with respect to $q_{1|1}$. That is, let

$$\begin{aligned} g(q_{1|1}) = \max_{q_{1|0}} \quad & \frac{p_1 - q_{1|0}}{q_{1|1} - q_{1|0}} \\ \text{s.t.} \quad & p_{1|a} = q_{1|0}(1 - \pi_{1|a}) + q_{1|1}\pi_{1|a} \ \ \forall a \\ & 0 \leq \pi_{1|a} \leq 1 \ \ \forall a \\ & 0 \leq q_{1|0} < q_{1|1} \end{aligned}$$

In this case, all constraints are satisfied so long as $0 \leq q_{1|0} \leq \min_a p_{1|a}$ and the maximum is achieved when $q_{1|0} = 0$. Thus, $g(q_{1|1}) = \frac{p_1}{q_{1|1}}$. Next, we solve

$$\begin{aligned} \max_{q_{1|1}} \quad & g(q_{1|1}) = \frac{p_1}{q_{1|1}} \\ \text{s.t.} \quad & p_{1|a} = q_{1|1}\pi_{1|a} \ \ \forall a \\ & 0 \leq \pi_{1|a} \leq 1 \ \ \forall a \end{aligned}$$

In this case, all constraints are satisfied so long as $\max_a p_{1|a} \leq q_{1|1} \leq 1$ and the maximum value that satisfies this constraint is $\frac{p_1}{\max_a p_{1|a}}$. Applying similar reasoning to the minimization problem, we get a minimum value of $\frac{p_1 - \min_a p_{1|a}}{1 - \min_a p_{1|a}}$. Thus, when $q_{1|1} > q_{1|0}$, we have the following sharp bounds on π_1

$$\frac{p_1 - \min_a p_{1|a}}{1 - \min_a p_{1|a}} \leq \pi_1 \leq \frac{p_1}{\max_a p_{1|a}} \tag{4}$$

Finally, we repeat this derivation for $q_{1|1} < q_{1|0}$ and take the union of these two sets of bounds to get the bounds in Proposition 4.

4 EXAMPLE LINEAR PROGRAMS

In this section, we present example linear programs for each of the general IV models presented in Figure 1 in Section 4 of the main paper. In all cases, we present bounds for $\mathbb{E}[X]$

4.1 FIGURE 1 (B)

In this model, A can be used to represent observed randomness in the measurement method. Because, X has no parents, $\tilde{X} = X$, and thus $\psi = P(X, \tilde{Y})$

$$\begin{aligned} \text{objective:} \ & \sum_{a,x,\tilde{y}} xP(A=a)\psi_{x,\tilde{y}} \\ \text{s.t.} \ & \sum_{x,\tilde{y}} \mathbb{I}(\tilde{y}(x,a) = y)\psi_{x,\tilde{y}} = P(Y = y \mid A = a) \\ & \psi_{x,\tilde{y}} \geq 0. \end{aligned} \tag{5}$$

4.2 FIGURE 1 (C)

In this model, A and A' represent independent IVs.

$$\begin{aligned} \text{objective:} \ & \sum_{a,a',\tilde{x},\tilde{y}} x(a,a')P(A=a, A'=a')\psi_{\tilde{x},\tilde{y}} \\ \text{s.t.} \ & \sum_{\tilde{x},\tilde{y}} \mathbb{I}(\tilde{y}(x) = y, \tilde{x}(a,a') = x)\psi_{\tilde{x},\tilde{y}} = P(Y = y \mid A = a, A' = a') \\ & \psi_{\tilde{x},\tilde{y}} \geq 0. \end{aligned} \tag{6}$$

4.3 FIGURE 1 (D) AND (E)

After randomizing the instruments according to Procedure 1, the graphs for these models are identical to the classic IV model. Thus, the linear program for the classic IV model, shown in Equation 6, can be used to bound parameters of $P(A, X, Y)$.

References

Rudolf Beran. Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, 85(410): 417–426, 1990.

Rafael Chaves, Daniel Cavalcanti, and Leandro Aolita. Causal hierarchy of multipartite Bell nonlocality. *Quantum*, 1:23, August 2017. ISSN 2521-327X.

Jing Cheng and Dylan S. Small. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(5):815–836, 2006.

Vasek Chvatal. *Linear Programming*. W.H. Freeman, 1983.

Robin J. Evans. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

Robin J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 2016.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *Review of Economic Studies*, 86(1): 377–412, 2019.

Arthur Fine. Hidden variables, joint probability, and the bell inequalities. *Physics Review Letters*, 48, 1982.

Aurélien Garivier. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *In Proceedings of COLT*, 2011.

D Gieger and Christopher Meek. Quantifier elimination for statistical problems. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

Joel L. Horowitz and Charles F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84, 2000.

Robert Nowak and Ervin Tánczos. Tighter confidence intervals for rating systems, 2019.