Partial identification without distributional assumptions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal effect estimation is important for numerous tasks in the natural and social sciences. However, identifying effects is impossible from observational data without making strong, often untestable assumptions which might not be applicable to real-world data. We consider algorithms for the *partial* identification problem, bounding the effects of multivariate, continuous treatments over multiple possible causal models when unmeasured confounding makes identification impossible. Even in the partial identification setting, most current work is only applicable in the discrete setting. We propose a framework which is applicable to continuous high-dimensional data. The observable evidence is matched to the implications of constraints encoded in a causal model by norm-based criteria. In particular, for the IV setting, we present ways by which such constrained optimization problems can be parameterized without likelihood functions for the causal or the observed data model, reducing the computational and statistical complexity of the task.

1 Introduction

Estimating causal effects is a key goal of scientific inquiry, enabling better decision making in medicine [5], economics [6], epidemiology [23], and beyond [20, 14]. The gold standard for estimating effects is the randomized controlled trial (RCT). RCTs work because randomization removes *confounding*. However, in many cases it is physically, logistically, or ethically impossible to conduct RCTs. In such cases, causal effects can only be identified under structural assumptions [20, 12]. When unobservable confounders are present, techniques such as instrumental variable (IV) models [12, Ch. 16], the front-door criterion [20, Ch. 3] and proxies [22] may be applicable. Even then, identification requires further assumptions, not just about the graphical structure but about the functional form of causal associations, e.g., monotonicity [1], additivity [11, 21, 17, 4] or conditions such as completeness [22], which are not immediately intuitive as they refer to unspecified hidden variables, and likely do not apply to most real-world scenarios. When such assumptions fail, effects may still be *partially identifiable*—i.e., bounded with respect to the set of models consistent with the data [16]. This task has not received nearly as much attention as the point estimation problem, despite promising early work in this area [7, 3]. Lately the topic has sparked some interest. Xia et al. [24] outline a procedure for bounding causal effects with neural networks but restrict their focus to discrete, low-dimensional data. Duarte et al. [8], Zhang et al. [25] reduce the bounding problem to a polynomial program, but both assume that variables are discrete and [25] further assumes that variables are finite. Kilbertus et al. [15] propose a method for computing causal bounds in IV models with continuous treatments using gradient descent and Monte Carlo integration. Their procedure is limited to univariate settings. We extend this work to higher dimensions and different causal structures, replacing the copula formulation of Kilbertus et al. [15] with a more generic parameterization that decouples model fitting from changes due to the unobserved confounder U.

We make the following contributions: (1) We propose a generic, modular form of the effect bounding problem compatible with a wide range of graphical structures, function classes, and distance measures,

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

(2) We derive an efficient solution to the problem via constraint sampling and automatic differentiation, instantiated for the Instrument variable (IV). Critically, we avoid full density matching, instead only matching the first two moments of a single conditional distribution. (3) We illustrate our method on synthetic and semi-synthetic datasets, where simulations confirm that our approach computes valid bounds even in complex settings where common assumptions fail. While we do introduce a general framework, we focus on 'coarse' causal graphs, since they are the most useful in practice. The goal is to allow for a real-world domain expert to be able to deliver results with minimal assumptions about non-causal aspects of the model.

2 Setup

We first formulate the general setup and then provide a concrete example based on the IV model setting. Assume we observe data from causal model $S^\star$, with density function $p_{S^\star}$ (of which we have an estimate $\hat{p}$), including a continuous treatment $X \in \mathbb{R}^p$ and outcome $Y \in \mathbb{R}$ among the observed variables. Using Pearl's *do* notation [20], the estimand of interest is $o_{x^\star}(S^\star) := \mathbb{E}_{S^\star}[Y \mid do(X = x^\star)]$, typically not identified. Our goal will be to minimize/maximize $o_{x^\star}(S^\star)$ over all *admissible* causal models S among a (uncountable) model class $\mathcal{S}$. We characterize admissibility via two types of constraints. First, the missing edges of the assumed directed acyclic graph (DAG) encode conditional independencies between variables, which we call *structural constraints*. Second, we have the *data constraint* via a distance function $\mathrm{dist}(p_S, \hat{p})$ into $\mathbb{R}^+$ that measures the (estimated) discrepancy between model $S \in \mathcal{S}$ and the ground truth model $S^\star$.

Hence, we arrive at the following general problem setting for computing the minimum/maximum causal effect among all models $\mathcal{S}$ that are $(\mathrm{dist}, \epsilon)$-compatible with the observed data:

$$\min_{S \in \mathcal{S}} \,/\, \max \quad o_{x^\star}(S) \qquad \texttt{[obj]} \tag{1}$$

$$\text{subject to} \quad \mathrm{dist}(p_S, \hat{p}) \leq \epsilon\,, \qquad \texttt{[c-data]} \tag{2}$$

$$\text{structural constraints} \qquad \texttt{[c-struct]} \tag{3}$$

Consider the **instrumental variable** model in Figure 1(i). Following the *structural causal model* (SCM) notation of Pearl [20], we observe instruments $Z \in \mathcal{Z} \subset \mathbb{R}^q$, treatments $X \in \mathcal{X} \subset \mathbb{R}^p$, and outcomes $Y \in \mathcal{Y} \subset \mathbb{R}$ with a potentially high-dimensional confounded pair of *background* variables U_X and U_Y. Here we assume (A1) $X \not\perp\!\!\!\perp Z$, (A2) $Z \perp\!\!\!\perp \{U_X, U_Y\}$, and (A3) $Z \perp\!\!\!\perp Y \mid \{X, U_X, U_Y\}$. We denote the i.i.d. observations that make up $\hat{p}$ by $\mathcal{D} = \{x_i, y_i, z_i\}_{i=1}^n$. Instead of the common additive noise or linearity assumptions, we consider a larger class of potentially *non-linear*, *non-additive* functions $f : \mathcal{X} \times \mathcal{U}_Y \to \mathcal{Y}, g : \mathcal{Z} \times \mathcal{U}_X \to \mathcal{X}$ that we will later encode in $\mathcal{S}$. We propose to avoid full density estimation for `[c-data]` and instead only estimate $X \mid Z$ and the first two moments of $Y \mid \{X, Z\}$. This is to say that $\mathrm{dist}(p_S, \hat{p})$ here is the matching of the first two moments of $Y \mid \{X, Z\}$.

3 Method

In the following, we assume the IV setting and describe the method at a high level for brevity.

The Response function framework [2, 15] allows us to sidestep the complexity of dealing with the potentially infinite dimensional confounders without loss of generality. We call $f_u(\cdot) := f(\cdot, u) : \mathcal{X} \to \mathcal{Y}$ the response function for the fixed value $U = u$ of the background variable U. A distribution over values of U entails a distribution over response functions f_u. Hence, instead of modeling U and f separately as part of a causal model S, we can directly encode them jointly via a distribution over response functions. Without restrictions on the functional dependence of X and Y on U, one cannot obtain non-trivial bounds [9, 10]. Specifically, we choose a family of response functions $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$ that captures plausible direct causal effects from X to Y under all possible values of the confounding variable(s). We then assume a parameterized family of distributions over $\mathcal{F}$, denoted by $\{p_\eta^{\mathcal{F}} \mid \eta \in \mathbb{R}^d\}$. We propose parameterized response functions as linear combinations of a set of non-linear basis functions $\{\psi_k : \mathcal{X} \to \mathcal{Y}\}_{k=1}^K$ for some $K \in \mathbb{N}$:

$$\mathcal{F} := \left\{ f_\theta := \sum_{k=1}^K \theta_k \psi_k \,\middle|\, \theta_k \in \mathbb{R} \right\}. \tag{4}$$

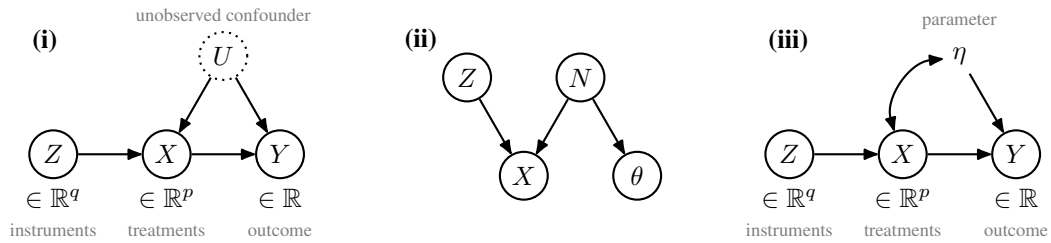

Figure 1: **(i)** The instrumental variable setting.**(ii)** The graphical model encoding, i.e., a parameterized generative distribution $p_\eta(\theta)$ that satisfies `[c-struct]` and matches the observed conditional $\hat{p}(x \mid z)$. **(iii)** The response function framework in which we express possible distributions of the confounder U and functions g, f with a distribution over functions $\mathcal{X} \to \mathcal{Y}$ parameterized by η.

Hence, once $\mathcal{F}$ is fixed, $p_\eta^{\mathcal{F}}$ is fully described as a distribution over $\theta \in \mathbb{R}^K$. For ease of notation, we summarize $\psi := (\psi_1, \ldots, \psi_K)$. The choice of basis allows the practitioner a continuum of choices regarding assumptions about the functional relationship between X and Y, linear or not.

3.1 Parameterizing $\mathcal{S}$ and satisfying `[c-struct]`

Next, we describe how to leverage the response function framework for a useful parameterization of causal models $\mathcal{S}$. For clarity, $\overline{\mathcal{S}} := \{p_\eta \mid \eta \in \mathbb{R}^d; \text{(A1)-(A3) satisfied for } (X, Y, Z) \sim p_\eta\}$.

Using the parameterization from Equation (4), we may now reformulate (A2) and (A3) as $Z \perp\!\!\!\perp \theta$ and $Z \not\perp\!\!\!\perp \theta \mid X$. Consequently, we propose the DAG in Figure 1(ii) as a model for θ, encoding $\overline{\mathcal{S}}$. The main takeaways here are

(i) The model encodes (A1), (A2) and (A3). (ii) A model for $X \mid \{Z, N\}$ can be fit once upfront from observed data and can be chosen as any invertible conditional model $h_Z(N)$ for which $h^{-1}_{Z=z}(x)$ exists uniquely for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$ (implying $d_N = p$). We use conditional normalizing flows[19], but the practitioner has a free choice here.[1].

We model θ with a *partially specified* multivariate distribution as follows,

$$\theta \mid N \sim p_{\eta.\theta}\big(\cdot; \mu_{\eta_0}(N), \Sigma_{\eta_1}(N)\big), \tag{5}$$

with mean and covariance functions $\mu_{\eta_0}(N)$ and $\Sigma_{\eta_1}(N)$, parameterized by neural networks with parameters η_0 and η_1 in our experiments. The combined parameters $\eta = (\eta_0, \eta_1)$ now encode our family of causal models $\overline{\mathcal{S}} = \{p_\eta \mid \eta \in \mathbb{R}^d\}$, where Y is implicitly given by the random variable $f_\theta(X)$ with $X = h_Z(N)$, $N \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and the required moments of $\theta \mid N$ as in Equation (5).

We now rephrase our main optimization problem as:

$$\min_{\eta \in \mathbb{R}^d} \,/\, \max \quad o_{x^\star}(\eta) = \psi(x^\star)^\top \, \mathbb{E}_N[\mu_{\eta_0}(N)] \qquad \texttt{[obj]}$$

$$\text{subject to} \quad \operatorname{dist}(p_\eta, \hat{p}) \leq \epsilon \,. \qquad \texttt{[c-data]}$$

For the objective equation, we used the law of total expectation. Since N follows a standard Gaussian, we can easily estimate the remaining expectation from a finite sample. So far, the data have only entered once when we chose $\overline{\mathcal{S}}$ such that $\hat{p}(X \mid Z)$ is matched by all $p_\eta \in \overline{\mathcal{S}}$. We still must enforce the remaining data matching constraints `[c-data]`, which amounts to defining dist.

3.2 Satisfying `[c-data]`: matching the data

We first factorize $p(Z, X, Y) = p(Y \mid X, Z)\, p(X \mid Z)\, p(Z)$. A key advantage of our construction of $\overline{\mathcal{S}}$ is that we only need to match the factor $p(Y \mid X, Z)$ to satisfy `[c-data]` instead of having to perform density estimation to match high-dimensional distributions via, e.g., f-divergences, integral probability metrics, Stein discrepancy, adversarial optimization, or variational inference. This matching can be conveniently formulated as matching expectations of the two distributions when transformed by a fixed set of dictionary functions $\{\phi_l : \mathcal{Y} \to \mathbb{R}\}_{l=1}^L$. That is, we want the absolute values of

$$\nu_l(x, z) = \mathbb{E}_{y \sim \hat{p}(y \mid x, z)}[\phi_l(y)] - \mathbb{E}_{\theta \sim p_\eta(\theta \mid x, z)}[\phi_l(f_\theta(x))] \tag{6}$$

[1] N is merely a device for parameterizing $p(x \mid z)$ and $p(\theta \mid x, z)$, and is not meant to have a causal interpretation

to be small for all l and all $(x, z) \in \mathcal{X} \times \mathcal{Z}$. To simplify optimization and modeling, in practice it often suffices to constrain the first few moments of $Y \mid \{X, Z\}$, e.g., to set $\phi_1(Y) = Y$ and $\phi_2(Y) = Y^2$, which is what we do. This task is much more data efficient than modeling the complete joint distribution $p(X, Y, Z)$, as required, for example, by generative methods such as GANs [13]. Since X, Z are continuous (and potentially high-dimensional), this would amount to an uncountably infinite set of constraints. In high dimensions, a naïve discretization of X, Z also fails [15]. Instead, we aim to make $|\nu_l(x, z)|$ small at a representative, finite set of points in $\mathcal{X} \times \mathcal{Z}$. Arguably, the most representative set is a uniformly random subsample of the observed data $\mathcal{D}$ of size M. For notational simplicity, we assume w.l.o.g. that these "support points" are the first M indices. In other words, defining (by some overloading of notation) the matrix $\nu \in \mathbb{R}^{L \times M}$ via $\nu_{l,j} = \nu_l(x_j, z_j)$ (using $j = 1, \ldots, M$ to index the selected support points), we obtain a natural family of overall distances $\operatorname{dist}(p_\eta, \hat{p}) = \|\nu\|$ for any (semi-)norm $\|\cdot\|$ on $\mathbb{R}^{L \times M}$. In our experiments, we specifically consider the entry-wise sup-norm $\|\cdot\|_{\infty,\infty}$ and the entry-wise 2-norm $\|\cdot\|_{2,2}$. For $\|\cdot\|_{\infty,\infty}$ we require the absolute value of each entry to be small, i.e., practically have to enforce $M \cdot L$ constraints, whereas for $\|\cdot\|_{2,2}$ we get away with a single overall constraint. Depending on the problem at hand, the practitioner can choose a larger value of l (consider more moments), choose an entirely different distance function, or choose norms different to the ones we choose. This again highlights the flexible and modular nature of our approach and the continuum of assumptions which the framework easily allows us to make.

For our assumptions, the **first expectation** in Equation (6) is estimable from the observed data via $\hat{\phi}_l(x, z) \approx \mathbb{E}[\phi_l(Y) \mid x, z]$, where $\hat{\phi}_l : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ are regression functions mapping $x, z \to \mathbb{E}[\phi_l(Y) \mid x, z]$ that were learned in a supervised fashion on the training data $\mathcal{D}$. Because h_Z was assumed to be invertible, our modeling choice for $\overline{\mathcal{S}}$ plays nicely with our constraint formulation, in that we can explicitly write

$$\theta \mid x, z \sim \mathcal{F}_\theta\Big(\theta; \mu_{\eta_0}\big(h_z^{-1}(x)\big), \Sigma_{\eta_1}\big(h_z^{-1}(x)\big)\Big) \,. \tag{7}$$

For general dictionary functions ϕ_l, we can again estimate the **second expectation** in Equation (6) using finite samples from a fully-specified version of Equation (7). This further simplifies the second expectation in Equation (6) to the following:

$$A_{1,j}(\eta) = \psi(x_j)^\top \mu_{\eta_0}(n_j) \,, \quad A_{2,j}(\eta) = \psi(x_j)^\top \big(\Sigma_{\eta_1}(n_j) + \mu_{\eta_0}(n_j)\mu_{\eta_0}(n_j)^\top\big)\psi(x_j) \,, \tag{8}$$

where $n_j := h_{z_j}^{-1}(x_j)$ for $j \in [M]$.

3.3 Solving the optimization

Taking all the steps from previous sections together, we have $\nu_{l,j} = \hat{\phi}_l(x_j, z_j) - A_{l,j}(\eta)$ and ultimately arrive at the following non-convex optimization problem with non-convex constraints

$$\begin{aligned} &\min_{\eta \in \mathbb{R}^d} / \max \; o_{x^\star}(\eta) = \psi(x^\star)^\top \, \mathbb{E}_N[\mu_{\eta_0}(N)] && \texttt{[obj]} \\ &\text{subject to } \|\nu\| \leq \epsilon && \texttt{[c-data]} \end{aligned} \tag{9}$$

for which the augmented Lagrangian method with inequality constraints is a natural choice [18, Sect. 17.4]. We provide a high-level description in Algorithm 1. Note that for the entry-wise sup-norm $\|\cdot\|_{\infty,\infty}$ we aim at enforcing $M \cdot L$ inequality constraints, one for each entry.

4 Experiments

Since ground truth causal effects must be known to properly evaluate the validity of our bounds, we make use of simulated datasets in a partially identifiable setting.

4.1 Treatment choices

When visualizing results for multidimensional treatments X, we vary the interventional values $x^\star$ along a single treatment dimension, keeping the remaining components at fixed values. While this allows us to show continuous treatment effect curves, we note our method can compute bounds for any multidimensional intervention $do(X = x^\star)$. Specifically, in the results shown here, we vary the first component of X and fix the values of the other components to their empirical marginal means. For each figure, we include a kernel density estimate of the marginal distribution of the empirically observed treatments to distinguish "data poor" from "data rich" regions. In "data poor" regions, we

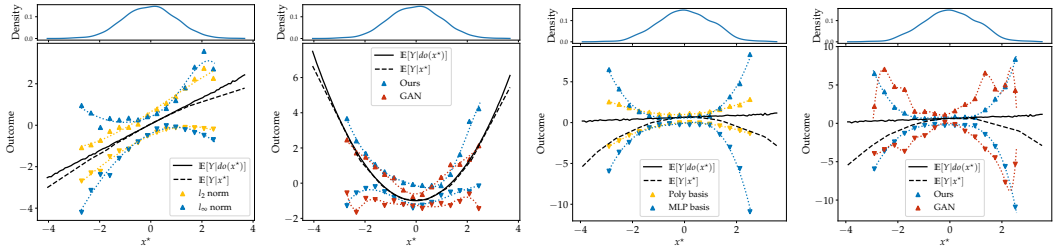

Figure 2: **Left** (lin-2d-weak; comparing norms) Our framework easily allows for different data-matching criteria (dist). The choices shown here ($\|\cdot\|_{\infty,\infty}$ and $\|\cdot\|_{2,2}$) yield comparable and consistent bounds. **Second Left** (quad-2d-weak; quadratic f) Our bounds are always valid, while the GAN framework gives some invalid bounds **Second Right** (lin-2d-strong; comparing response functions) As expected, neural response functions give wider bounds than a linear polynomial basis because of being less expressive. **Right** (lin-2d-strong; GAN comparison) Our method reliably yields valid bounds even under strong confounding, while the GAN framework becomes unstable, potentially due to difficulties with the bilevel optimization problem.

may expect our bounds to become looser, as less information about the data-matching constraints is available. **Baselines.** We compare our method to a naïve regression and the method of Hu et al. [13]. We are unaware of any other competitor capable of estimating bounds in the presence of multivariate, continuous treatments/instruments/mediators.

- **Regression with MLP.** We naïvely fit an MLP with quadratic loss to predict outcome Y from the multidimensional treatment X, a modeling approach that assumes no confounding at all.
- **GAN framework.** Hu et al. [13] parameterize the causal model (i.e., its exogenous random variables and structural equations) with neural networks and apply the adversarial learning framework to search the parameter space. We adjust the model in their code to our examples, but otherwise leave hyperparameter choices and convergence criteria untouched.

4.2 Implementation choices

Choice of response functions. As described in Equation (4), we choose linear combinations of non-linear basis functions $\{\psi_k\}_{k\in[K]}$. For our experiments we mostly work with a set of K *neural basis functions*, obtained from the last hidden-layer activations of an MLP fit to the observed data $\{(x_i, y_i)\}_{i\in[n]}$, as well as (multivariate) polynomials up to a fixed degree. For two-dimensional treatments, we use $K = 6$ and $K = 3$ for polynomial and neural basis functions, respectively. For three-dimensional treatments we use $K = 10$. For our method, we individually compute **lower** ($\bigtriangledown$) and **upper** ($\bigtriangleup$) bounds at multiple values $x^\star \in \mathbb{R}$ for one dimension of the treatment and show how it compares to the **true causal effect** (**——**) and **naïve regression** (**- - -**). The lines shown for **lower** and **upper** (**·····**) bounds are univariate cubic splines fit to the bounds for individual $x^\star$-grid values. We use $n = 10,000$ i.i.d. sampled datapoints for each experiment and subsample $M = 100$ datapoints uniformly at random for the data constraints `[c-data]`. For the $X \mid Z$ model in the IV setting we use a quadratic conditional spline normalizing flow.

Getting final bound estimates. For each $x^\star$, we run 5 optimizations with different seeds each for the lower and upper bounds. For the final upper bound estimate, we take the maximum of the 5 upper bound estimates we get, and for the final lower bound estimate we take the minimum of the 5 lower bound estimates we get. We follow the same process for the GAN bounds.

4.3 Results

We focus on datasets where the true effect is a linear or quadratic polynomial function of the treatment X. A glossary of datasets can be found in Appendix B, along with exact structural equations for each setting. Figure 3 provides a short description of the naming logic.

lin-2d-strong. This dataset is simulated from an IV setting, where the true effect $Y \mid do(X)$ is linear in a two-dimensional treatment X. The naïve regression $\mathbb{E}[Y \mid X]$ differs substantially from the true effect, indicating strong confounding (see, Figure 2 (Second right)). We start by assessing the effect of the flexibility of response functions on our bounds in Figure 2 (bottom left). Choosing less flexible basis functions (polynomials) yields tighter bounds, *highlighting the flexibility of our approach in obtaining more informative bounds when more restrictive assumptions are made*. Our method can

also accommodate alternative constraint formulations and slack parameters. While sometimes loose, especially for flexible basis functions (MLP) and in "data poor" regions towards the tails of the empirical distribution of observed treatments, our bounds contain the true causal effect for all $x^\star$. This validity holds for all other experiments as well. Next, Figure 2 (Right) compares our approach to the GAN framework baseline. First, we note that methods that fit a model and bounds at the same time, like Hu et al. [13], essentially need to solve a bilevel optimization problem, where the same parameters are shared in the definition of the constraints (which require optimizing a measure of fitness) and the causal quantity of interest (which requires optimizing a different problem). Such a strategy can be potentially more prone to optimization and regularization challenges. Our goal in comparing[2] to Hu et al. [13] is to show that our framework simplifies constraints through structural encoding and avoids full density estimation while giving comparable or better results than those from the GAN framework. Indeed, Figure 2 (bottom right) shows that despite the flexible neural basis functions, our approach can yield tighter bounds and avoid the instabilities observed in the GAN approach when we move towards the tails of the observational distribution.

lin-2d-weak. This is an IV setting with weak confounding. In Figure 2 (Left) we show how our bounds behave under different choices of dist, namely under the entry-wise $\|\cdot\|_{\infty,\infty}$ norm (which results in $M \cdot L$ constraints in the augmented Lagrangian) versus the entry-wise $\|\cdot\|_{2,2}$ norm (yielding only a single constraint). The obtained bounds are compatible and comparable, indicating relatively mild effects of the choice of the data-matching criterion dist.

quad-2d-weak. This is an IV setting with weak confounding and quadratic true effect. We do note that in this instance, the bounds given by our method are not as smooth in the data poor regions. However, our bounds are always valid, while the GAN framework gives some invalid bounds

5 Discussion

Limitations. After parameterization, the optimization problem in Equation (1) will generally result in a *non-convex objective with non-convex constraints*. Therefore, our proposed gradient-based local optimization may not converge to a global optimum, possibly rendering our bounds overly tight. Empirically, we do not observe evidence of consistently getting stuck in bad local optima. Because we optimize both bounds individually for each value of $x^\star$, our method may be computationally expensive when the intervention space is very large. However, it is well-suited for scenarios where we want reliable bounds on a well-defined set of plausible interventions. Finally, we have not accounted for the uncertainty of our bounds. Confidence or credible intervals for both extrema can help practitioners evaluate the reliability of causal inferences, and are an interesting direction for future work.

Conclusion. Causal modeling inevitably involves a trade-off between the strength of input assumptions and the specificity of resulting inferences. If assumptions do not correspond to the real world, the inference that follows is unwarranted. Our framework allows for a real-world domain expert to focus more on the coarse structure of the problem and assumptions about the function space, in contrast to a complex causal graph that is too prone to misspecification, and to deliver results with minimal assumptions about non-causal aspects of the model, while also having the flexibility of making stronger assumptions when appropriate.

We have introduced a stochastic causal program for bounding treatment effects in partially identifiable settings. Our approach does not rely on the typical assumptions of linearity, monotonocity, or additivity. We presented a novel parameterization for decoupling observed from unobserved potential outcomes, and derived an efficient procedure for optimizing over a set of causal models consistent with the data. Experiments demonstrate that our method produces valid and informative bounds in a wide range of settings, including with continuous and multidimensional variables.

While only the IV setting has been presented here, the method can be more widely applicable, and has already been applied to a leaky mediator setting in a longer version of the paper. The next steps would be to apply the method to a real world dataset. Causal conclusions of real world implications deserve thinking and exposition that are not easily summarized by a non-synthetic benchmark, and one of our goals in wanting to present in this workshop is to find interested experts to do exactly that.

[2]For generality, we use the *neural basis functions* for all experiments where we compare to the GAN framework, as they are expected to give wider bounds than polynomials.

References

[1] Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.*, 90(430):431–442, 1995.

[2] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 46–54, 1994.

[3] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.*, 92(439):1171–1176, 1997.

[4] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, 2019.

[5] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela van der Schaar. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.*, 109(1), 2021.

[6] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud.*, 86(5):1867–1900, 2018.

[7] David Maxwell Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pages 1269–1276, 1996.

[8] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *arXiv* preprint, 2109.13471, 2021.

[9] F F Gunsilius. Nontestability of instrument validity under continuous treatments. *Biometrika*, 108(4):989–995, 2021.

[10] Florian Gunsilius. Bounds in continuous instrumental variable models. *arXiv* preprint, 1910.09502, 2019.

[11] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[12] Miguel A. Hernán and James M. Robins. *Causal Inference: What If.* Chapman & Hall/CRC Studies, Boca Raton, 2020.

[13] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[14] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015.

[15] Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, 2020.

[16] Charles F Manski. Nonparametric bounds on treatment effects. *Am. Econ. Rev.*, 80(2):319–323, 1990.

[17] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 2710–2721, 2020.

[18] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[19] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Algorithm 1 Computing upper or lower bounds on $\mathbb{E}[Y \mid do(X = x^\star)]$ in the IV setting.

Require: dataset $\mathcal{D} = \{(z_i, x_i, y_i)\}_{i=1}^n$; constraint functions $\{\phi_l : \mathcal{Y} \to \mathbb{R}\}_{l=1}^L$; basis functions $\{\psi_k : \mathcal{X} \to \mathcal{Y}\}_{k=1}^K$; norm $\|\cdot\|$ for dist; batchsize for Monte Carlo B; number of support points M; tolerance $\epsilon > 0$

Setup: One-time computations shared for different $x^\star$ values

Fit (invertible) conditional normalizing flow $X = h_Z(N)$ from data $\mathcal{D}$ for $N \sim \mathcal{N}(0, \mathbf{1}_p)$

Fit MLPs $\hat{\phi}_1 : x_i, z_i \to y_i$ and $\hat{\phi}_2 : x_i, z_i \to y_i^2$ by minimizing the squared loss from data $\mathcal{D}$

subsample M indices from $[n]$ (uniform, no replacement) ▷ "support points", w.l.o.g. use $[M]$

Optimization: performed separately for lower and upper bound for each $x^\star$

minimize OBJECTIVE(η) subject to CONSTRAINT(η) $\leq \epsilon$

function OBJECTIVE(η)

  $o_{x^\star}(\eta) \leftarrow \psi(x^\star)^\top \frac{1}{B} \sum_{j=1}^B \mu_{\eta_0}(n_j)$ with $n_j \sim \mathcal{N}(0, \mathbf{1}_p)$ ▷ differentiable w.r.t. η

  return $\pm o_{x^\star}(\eta)$ ▷ objective, $\pm$ for lower/upper bound

function CONSTRAINT(η)

  $n_j \leftarrow h_{z_j}^{-1}(x_j)$ for $j \in [M]$ ▷ invert $X \mid Z$ model to infer "noises"

  $A_{1,j}(\eta) \leftarrow \psi(x_j)^\top \mu_{\eta_0}(n_j)$ for $j \in [M]$ ▷ moments implied by model

  $A_{2,j}(\eta) \leftarrow \psi(x_j)^\top \Big(\Sigma_{\eta_1}(n_j) + \mu_{\eta_0}(n_j)\mu_{\eta_0}(n_j)^\top\Big)\psi(x_j)$ for $j \in [M]$

  $\nu_{l,j} \leftarrow \hat{\phi}_l(x_j, z_j) - A_{l,j}(\eta)$ for $l \in \{1, 2\}, j \in [M]$ ▷ constraint matrix

  return $\|\nu\|$

[20] Judea Pearl. *Causality*. Cambridge University Press, New York, 2009.

[21] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

[22] Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv:2009.10982*, 2022.

[23] Jan P Vandenbroucke, Alex Broadbent, and Neil Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int. J. Epidemiol.*, 45(6):1776–1786, 2016.

[24] Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The neural-causal connection: expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems*, 2021.

[25] Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. *arXiv* preprint, 2110.05690, 2021.

A Pseudocode

Algorithm 1 showcases a possible implementation of the partial identification procedure.

B Glossary of Synthetic Datasets

We describe here the synthetic datasets we use in our experiments.

lin - 2d - strong

Figure 3: Naming logic for the datasets. The *first* segment says whether y is linear or quadratic in x. The *second* segment tells the dimension of the treatment. The *third* segment denotes the strength of the confounding, strong or weak.

We use various polynomial datasets where the causal effect is a polynomial function of a single or multidimensional treatment. We provide the construction of each of these here. Every node except

Y (outcome) could be multi-dimensional. If a node, say C, is multi-dimensional, we index the dimension with a subscript. That is, we write $C = (c_1, c_2)$. Figure 3 visually describes the naming logic behind the datasets to make the exposition clearer.

The noises, confounder, and instruments follow

$$c, z, e_x \sim \mathcal{N}^2(0, 1)$$
$$e_y \sim \mathcal{N}(0, 1)$$

- `lin-2d-strong` (f_y linear in x, strong non-additive confounding)

$$f_x(z, c, e_x) = 0.5z + 2c + e_x$$
$$f_y(x, c, e_y) = x_1 + x_2 - 3(x_1 + x_2)(c_1 + c_2) + e_y$$

- `lin-2d-weak` (f_y linear in x, weak non-additive confounding)

$$f_x(z, c, e_x) = 2z + c + e_x$$
$$f_y(x, c, e_y) = 5x_1 + 6x_2 - x_1(c_1 + c_2) + e_y$$

- `IV-quad-2d-weak` (f_y quadratic in x, weak non-additive confounding)

$$2f_x(z, c, e_x) = 2z + c + e_x$$
$$f_y(x, c, e_y) = 5x_1^2 + 6x_2^2 - (x_1 + x_2)(c_1 + c_2) + e_y$$