

---

# A Class of Algorithms for General Instrumental Variable Models

---

**Niki Kilbertus\***  
Helmholtz AI

**Matt J. Kusner**  
University College London  
The Alan Turing Institute

**Ricardo Silva**  
University College London  
The Alan Turing Institute

## Abstract

Causal treatment effect estimation is a key problem that arises in a variety of real-world settings, from personalized medicine to governmental policy making. There has been a flurry of recent work in machine learning on estimating causal effects when one has access to an instrument. However, to achieve identifiability, they in general require one-size-fits-all assumptions such as an additive error model for the outcome. An alternative is partial identification, which provides bounds on the causal effect. Little exists in terms of bounding methods that can deal with the most general case, where the treatment itself can be continuous. Moreover, bounding methods generally do not allow for a continuum of assumptions on the shape of the causal effect that can smoothly trade off stronger background knowledge for more informative bounds. In this work, we provide a method for causal effect bounding in continuous distributions, leveraging recent advances in gradient-based methods for the optimization of computationally intractable objective functions. We demonstrate on a set of synthetic and real-world data that our bounds capture the causal effect when additive methods fail, providing a useful range of answers compatible with observation as opposed to relying on unwarranted structural assumptions.<sup>1</sup>

## 1 Introduction

Machine learning is becoming more and more prevalent in applications that inform actions to be taken in the physical world. To ensure robust and reliable performance, many settings require an understanding of the causal effects an action will have before it is taken. Often, the only available source of training data is observational, where the actions of interest were chosen by unknown criteria. One of the major obstacles to trustworthy causal effect estimation with observational data is the reliance on the strong, untestable assumption of *no unobserved confounding*. To avoid this, only in very specific settings (e.g., front-door adjustment, linear/additive instrumental variable regression) it is possible to allow for unobserved confounding and still identify the causal effect (Pearl, 2009). Outside of these settings, one can only hope to meaningfully bound the causal effect (Manski, 2007).

In many applications, we have one or few treatment variables  $X$  and one outcome variable  $Y$ . Nearly all existing approaches to obtain meaningful bounds on the causal effect of  $X$  on  $Y$  impose constraints on how observed variables are related, in order to mitigate the influence of unobserved confounders. One of the most useful structural constraints is the existence of an observable *instrumental variable* (IV): a variable  $Z$ , not caused by  $X$ , whose relationship with  $Y$  is entirely mediated by  $X$ , see Pearl (2009) for a graphical characterization. The existence of an IV can be used to derive upper (lower) bounds on causal effects of interest by maximizing (minimizing) those effects among all IV models compatible with the observable distribution. *In this work, we develop algorithms to compute these*

---

\*Majority of work done while at Max Planck Institute for Intelligent Systems and University of Cambridge.

<sup>1</sup>Code available at <https://github.com/nikikilbertus/general-iv-models>.

*bounds on causal effects over “all” IV models compatible with the data in a general continuous setting.* Crucially, the space of “all” models *cannot* be arbitrary, but it can be made very flexible. Instead of forcing a user to adopt a model space with hard constraints, we will allow for choice from a continuum of model spaces. Our approach rewards background knowledge with tighter bounds and it is not tied to an a priori inflexible choice, such as additivity or monotonicity. It avoids the adoption of unwarranted structural assumptions under the premise that they are needed due to the lack of ways of expressing more refined domain knowledge. The burden of the trade-off is put explicitly on the practitioner, as opposed to embracing possibly crude approximations due to the limitations of identification strategies.

Eliciting constraints that characterize “the models compatible with data” under a causal directed acyclic graph (DAG) for discrete variables is an active field of study, with contributions from the machine learning, algebraic statistics, economics, and quantum mechanics literature. This has provided complete characterizations of equality (Evans, 2019; Tian & Pearl, 2002) and inequality (Wolfe et al., 2019; Navascues & Wolfe, 2019) constraints. Enumerating all inequality constraints is in general super-exponential in the number of observed variables, even for discrete causal models. However, this line of work typically solves a harder problem than is strictly required for bounding causal effects: they provide symbolic constraints obtained by eliminating all hidden variables. While the pioneering work of Balke & Pearl (1994) in the discrete setting also provides symbolic constraints via a super-exponential algorithm, it introduces constraints that match the observed marginals of a latent variable model against the observable distribution. Thereby it provides a connection to non-symbolic, stochastic approaches for evaluating integrals, which we develop in this work.

Our key observation is that we can leverage recent advances in efficient gradient and Monte Carlo-based optimization of computationally intractable objective functions to bound the causal effect directly. This can be done even in the setting where  $X$  is continuous, where none of the literature described above applies. We do so by (a) parameterizing the space of causal responses to treatment  $X$  such that we can incorporate further assumptions that lead to informative bounds; (b) using a Monte Carlo approximation to the integral over the distribution of possible responses to  $X$ , where the distribution itself must be parameterized carefully to incorporate the structural constraints of an IV DAG model. This allows us to optimize over the domain-dependent set of all plausible models that are consistent with observed data to find lower/upper bounds on the target causal effect.

In Section 2, we describe the general problem of using instrumental variables when treatment  $X$  is continuous. Section 3 develops our representation of the causal model. In Section 4 we introduce a class of algorithms for solving the bounding problem and our suggested implementation. Section 5 provides several demonstrations of the advantages of our method.

## 2 Current Approaches and Their Limitations

Balke & Pearl (1994) focused on partial identification (bounding) of causal effects on binary discrete models. Angrist et al. (1996) studied identification of effects for a particular latent subclass of individuals also in the binary case. Meanwhile, the econometrics literature has focused on problems where the treatment  $X$  is continuous (Newey & Powell, 2003; Blundell et al., 2007; Angrist & Pischke, 2008; Wooldridge, 2010; Darolles et al., 2011; Horowitz, 2011; Chen & Christensen, 2018; Lewis & Syrgkanis, 2018). This problem has recently received attention in machine learning, using techniques from deep learning (Hartford et al., 2017; Bennett et al., 2019) and kernel machines (Singh et al., 2019; Muandet et al., 2020). This literature assumes that the structural equation for  $Y$  has a special form, such as having an additive error term  $e_Y$ , as in  $Y = f(X) + e_Y$ . The error term  $e_Y$  is not caused by  $X$ , but need not be independent of it, introducing unobserved confounding. This assumption is also used in related contexts, such as in sensitivity analysis for counterfactual estimands, see Kilbertus et al. (2019) for a specific application in fairness.

Using the notation of Pearl (2009), the expected response under an intervention on  $X$  at level  $x$  is denoted by  $\mathbb{E}[Y | do(x)]$ , which in the model above boils down to  $f(x)$ . An *average treatment effect* (ATE) can be defined as a contrast of this expected response under two treatment levels, e.g.,  $f(x) - f(x')$ . In the zero-mean additive error case,  $\mathbb{E}[Y | z] = \int f(x)p(x | z) dx$ . Under some regularity conditions, no function other than  $f(\cdot)$  satisfies that integral equation. Since  $\mathbb{E}[Y | z]$  and  $p(x | z)$  can be both learned from data, this allows us to learn the ATE from observational data. This

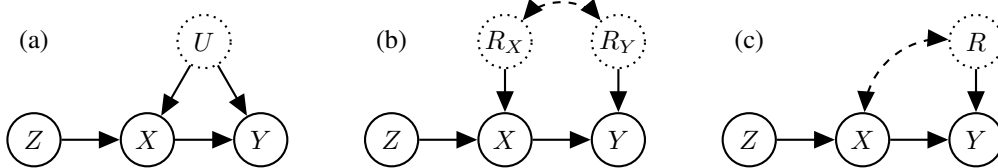


Figure 1: (a) An example of DAG compatible with  $Z$  being an instrument for  $X \rightarrow Y$ , with hidden confounder  $U$ . (b) An equivalent representation using *response function* indices for deterministic functions  $X = g_{R_X}(Z)$  and  $Y = f_{R_Y}(X)$ , with two random indexing variables  $R_X$  and  $R_Y$ . (c) For the purposes of modeling  $\mathbb{E}[Y | do(x)]$ , it is enough to express the model in terms of  $R := R_Y$  only.

is how the vast majority of recent work identifies the causal treatment effect in the IV model (Hartford et al., 2017; Bennett et al., 2019; Singh et al., 2019; Muandet et al., 2020).

The price paid for identification is that it seriously limits the applicability of these models. Diagnostic tests for the additivity assumption are not possible, as residuals  $Y - f(X)$  can be arbitrarily associated with  $X$  by assumption. On the other hand, without *any* restrictions on the structural equations, it is not only impossible to identify the causal effect of the IV model with a continuous treatment, but even bounds on the ATE are vacuous (Pearl, 1995; Bonet, 2001; Gunsilius, 2018, 2020). However, with relatively weak assumptions on the space of allowed structural equations, it is possible to achieve meaningful bounds on the causal effect (Gunsilius, 2020). It suffices that the equations for  $X$  and  $Y$  have a finite number of discontinuities. Gunsilius provides a theoretical framework for representation and estimation of bounds. Algorithmically, he proposes a truncated wavelet representation for the causal response and builds convex combinations of a sample of response functions to optimize IV bounds. Although it is an important proof of concept for the possibility of bounds for the general IV case with a strong theoretical motivation, we found that the method has frequent stability issues that are not easy to diagnose. We return to this in Appendix A.

Building on top of this work and some classical ideas first outlined by Balke & Pearl (1994), we propose an alternative formulation for finding bounds when both  $X$  and  $Y$  are continuous. Our technique flexibly parameterizes the causal response functions, while naturally encoding the structural IV constraints for compatibility with the observed data. We then leverage an augmented Lagrangian method that is tailored to non-convex optimization with inequality constraints. We demonstrate that our method matches estimation results of prior work in the additive setting, and gives meaningful bounds on the causal effect in general, non-additive models. Thereby, we follow a line of recent successes in various domains achieved by replacing previous intractable symbolic-combinatorial algorithms (Balke & Pearl, 1994; Wolfe et al., 2019; Drton et al., 2009) with a continuous program. One of our key contributions is to formulate bounds on true causal effects as well as their compatibility requirements as a smooth, constrained objective, for which we can leverage efficient gradient-based optimization techniques with Monte Carlo approximations.

### 3 Problem Setting

Following Pearl’s Structural Causal Model (SCM) framework (Pearl, 2009), we assume the existence of structural equations and a (possibly infinite dimensional) unobserved exogenous process  $U$ ,

$$X = g(Z, U) \quad \text{and} \quad Y = f(X, U). \quad (1)$$

We illustrate this situation in Figure 1(a). It assumes the usual requirements for the instrument  $Z$  to be satisfied, namely (a)  $Z \perp\!\!\!\perp U$ , (b)  $Z \not\perp\!\!\!\perp X$ , and (c)  $Z \perp\!\!\!\perp Y | \{X, U\}$ .

#### 3.1 Goal

The goal is to compute lower/upper bounds on  $\mathbb{E}[Y | do(x^*)]$  for any desired intervention level  $x^*$ . Bounds on (conditional) ATEs can be derived, see also Appendix B. Intuitively, we put bounds on how  $f(X, U)$  depends on  $X$  by optimizing over “allowed” distributions of  $U$ . Which distributions are “allowed” is determined by observations, i.e., we only consider settings where marginalizing  $U$  results in  $p(x, y | z)$  for all  $(x, y, z)$  in the support of the observational distribution. In fact, as pointed out by Palmer et al. (2011), it is enough to consider matching the marginals of the latent variable

model to the two conditional densities  $p(x|z)$  and  $p(y|z)$ <sup>2</sup>. Informally, *among all possible structural equations  $\{g, f\}$  and distributions over  $U$  that reproduce the estimated densities  $\{\hat{p}(x|z), \hat{p}(y|z)\}$ , we find estimates of the minimum and maximum expected outcomes under intervention.*

**Response functions.** The main idea of Balke & Pearl (1994) is to express structural equations in terms of *response functions*: labeling (and possibly clustering) states of  $U$  according to the implied functional relationship between the observed variable and its direct causes. These  $U$  states are mapped to a particular level of an index variable  $R$ . For instance, if  $Y = f(X, U) = \lambda_1 X + \lambda_2 X U_1 + U_2$ , a two-dimensional  $U$  space in a linear, non-additive outcome function, we have that  $f(x, u) = \lambda_1 x + \lambda_2 x$  for  $u_1 = 1, u_2 = 0$ . We can define an implicit arbitrary value  $r$  such that  $f_r(x) = \lambda_r x$ ,  $\lambda_r = \lambda_1 + \lambda_2$ , the value “ $r$ ” being an alias for  $(1, 0)$  in the space of the confounders. The advantage of this representation is that we can think of a distribution over  $R$  as a distribution over functions of  $X$  alone. Otherwise we would need to deal with interactions between  $U$  and  $X$  on top of a distribution over  $U$ , itself of unclear dimensionality. In contrast, the dimensionality of  $R$  is the one implied by the particular function space adopted. Gunsilius (2020) provides a more thorough discussion of the role of response functions corresponding to a possibly infinite-dimensional  $U$ . Figure 1(b) shows a graphical representation of a system parameterized by response function indices  $R_X$  and  $R_Y$ , with a bi-directed edge indicating possible association between the two. In what follows, as there will be no explicit need for  $R_X$ , the causal DAG corresponding to our counterfactual model is shown in Figure 1(c)<sup>3</sup>. This itself departs from Balke & Pearl (1994) and Gunsilius (2020), having the advantage of simplifying the optimization and not assuming counterfactuals for  $X$  (which will not exist if  $Z$  is not a cause of  $X$  but just confounded with it). Furthermore, focusing on  $\{p(x|z), p(y|z)\}$  instead of  $p(x, y|z)$  does not require simultaneous measurements of  $X$  and  $Y$  (Palmer et al., 2011), see Appendix G for the latter. Within this framework, we can rewrite the optimization over allowed distributions of  $U$  into an optimization over allowed distributions of response functions for  $Y$ .

Without restrictions on the function space, non-trivial inference is impossible (Pearl, 1995; Bonet, 2001; Gunsilius, 2018). In our proposed class of solutions, we will adopt a parametric response function space: each response type  $r$  corresponds to some parameter value  $\theta_r \in \Theta \subset \mathbb{R}^K$  for some finite  $K$ . We write  $f_r(x) := f_{\theta_r}(x)$ . Going forward, we will simply use  $\theta$  to denote a specific response type and drop the index  $r$ . While our method works for any differentiable  $f_\theta$ , we will focus on linear combinations of a set of basis functions  $\{\psi_k : \mathbb{R} \rightarrow \mathbb{R}\}_{k \in [K]}$ <sup>4</sup> with coefficients  $\theta \in \Theta$ :

$$f_\theta(x) := \sum_{k=1}^K \theta_k \psi_k(x). \quad (2)$$

We propose to optimize over distributions  $p_{\mathcal{M}}(\theta)$  of the response function parameters  $\theta$  in the unknown causal model  $\mathcal{M}$ , subject to the observed marginal of the model,  $\int p_{\mathcal{M}}(x, y|z, \theta) p_{\mathcal{M}}(\theta) d\theta$ , matching the corresponding (estimated) marginals  $p(y|z)$  and  $p(x|z)$ . Notice that  $\theta \perp\!\!\!\perp Z$  is implied by  $Z \perp\!\!\!\perp U$  in the original formulation in terms of exogenous variables  $U$ . We assume a parametric form for  $p_{\mathcal{M}}(\theta)$  via parameters  $\eta \in \mathbb{R}^d$ , denoted by  $p_\eta(\theta)$ . We propose to use function families for  $p_\eta(\theta)$  that allow for practically low-variance Monte-Carlo gradient estimation via the reparameterization trick (Kingma & Welling, 2014) to learn  $\eta$  — more in Section 3.2.

**Objective.** An upper bound for the expected outcome under intervention can be directly written as

$$\max_{\eta} \mathbb{E}[Y | do(x^*)] = \max_{\eta} \int f_\theta(x^*) p_\eta(\theta) d\theta. \quad (3)$$

A lower bound can be found analogously by the minimization problem. When optimizing eq. (3) constrained by  $p(y|z)$  and  $p(x|z)$  in the sequel, it will be necessary to define  $p_\eta(x, \theta|z)$ <sup>5</sup>. In particular,  $\int p_\eta(x, \theta|z) dx = p_\eta(\theta|z) = p_\eta(\theta)$ . The last equality will be enforced in the encoding of  $p_\eta(x, \theta|z)$ , as we need  $Z \perp\!\!\!\perp \theta$  even if  $Z \not\perp\!\!\!\perp \theta|X$ . This encoding is introduced in Section 3.2, which will also allow us to easily match the marginal  $p(x|z)$ . In Section 3.3, we construct constraints for the optimization so that the marginal of  $Y$  given  $Z$  in  $\mathcal{M}$  matches the model-free  $p(y|z)$ .

<sup>2</sup>In Appendix G, we discuss the case where we match  $p(y|x, z)$ , which can further tighten bounds with some computational advantages and disadvantages compared to  $p(y|z)$ .

<sup>3</sup>It is also possible to represent only  $R_X$  and drop  $R_Y$ . Zhang & Bareinboim (2020) do this in a way that provides a new view of the discrete treatment case.

<sup>4</sup>We use the notation  $[K] := \{1, \dots, K\}$  for  $K \in \mathbb{N}_{>0}$ .

<sup>5</sup>We abuse notation slightly by expanding the definition of  $\eta$  to simultaneously signify all parameters specifying this joint distribution, as well as individual parameters specific to certain factors of the joint.

### 3.2 Matching $p(x | z)$ and Enforcing $Z \perp\!\!\!\perp U$

Instead of formulating the criterion of preserving the observed marginal  $p(x | z)$  as a constraint in the optimization problem, we bake it directly into our model.<sup>6</sup> To accomplish that, we factor  $p_\eta(x, \theta | z)$  as  $p(x | z)p_\eta(\theta | x, z)$ . The first factor is identified from the observed data and we can thus force our model to match it. The second factor must be constructed so as to enforce marginal independence between  $\theta$  and  $Z$  (as required by  $Z \perp\!\!\!\perp U$ ). We achieve that by parameterizing it by a copula density  $c_\eta(\cdot)$  that takes univariate CDFs  $F(\cdot)$ , which uniquely define the distributions, as inputs,

$$p_\eta(\theta | x, z) := c_\eta(F(x | z), F_\eta(\theta_1), \dots, F_\eta(\theta_K)) \prod_{k=1}^K p_\eta(\theta_k). \quad (4)$$

Here we assume that each component  $\theta_k$  of  $\theta$  has a Gaussian marginal density with mean  $\mu_k$  and variance  $\sigma_k^2$ , i.e.,  $p_\eta(\theta_k) = \mathcal{N}(\theta_k; \mu_k, \sigma_k^2)$ . Moreover, assuming  $c_\eta$  is a multivariate Gaussian copula density requires a correlation matrix  $S \in \mathbb{R}^{(K+1) \times (K+1)}$  for which we only keep a Cholesky factor  $L$  without further constraints, rescaling  $L^\top L$  to have a diagonal of 1s. Our full set of parameters is

$$\eta := \{\mu_1, \ln(\sigma_1^2), \dots, \mu_K, \ln(\sigma_K^2), L\} \in \mathbb{R}^{K(K+1)/2+2K}.$$

### 3.3 Matching $p(y | z)$

In the continuous output case, our parameterization implies the following set of integral equations

$$\Pr(Y \leq y | Z = z) = \int \mathbf{1}(f_\theta(x) \leq y) p_\eta(x, \theta | z) dx d\theta, \quad (5)$$

for all  $y \in \mathcal{Y}, z \in \mathcal{Z}$ , the respective sample spaces of  $Y$  and  $Z$ , where  $\mathbf{1}(\cdot)$  is the indicator function. These constraints immediately introduce two difficulties. First, we have an infinite number of constraints to satisfy. Second, the right-hand side involves integrating non-continuous indicator functions, which poses a problem for smooth gradient-based optimization with respect to  $\eta$ .<sup>7</sup>

To circumvent these issues, we first choose a finite grid  $\{z^{(m)}\}_{m=1}^M \subset \mathcal{Z}$  of size  $M \in \mathbb{N}$ , instead of conditioning on all values in  $\mathcal{Z}$ . We compute  $z^{(m)}$  from a uniform grid on the CDF  $F_Z$  of  $Z$ , i.e.,  $z^{(m)} := F_Z^{-1}(m/M+1)$  for  $m \in [M]$ . Second, to avoid the integration of non-continuous indicator functions, we can express the constraints of eq. (5) in terms of expectations over a dictionary of  $L$  basis functions  $\{\phi_l\}_{l=1}^L$ . This leads to the following constraints for  $p(y | z)$ :

$$\mathbb{E}[\phi_l(Y) | z^{(m)}] = \int \phi_l(f_\theta(x)) p_\eta(x, \theta | z^{(m)}) dx d\theta \quad \text{for all } l \in [L], m \in [M]. \quad (6)$$

This idea borrows from mean embeddings, where one can reconstruct  $p(y | z)$  from an infinite dictionary sampled at infinitely many points in  $\mathcal{Z}$  (Singh et al., 2019). In this work, we choose an even simpler approach and only constrain moments like mean and variance  $\phi_1(Y) := \mathbb{E}[Y]$ ,  $\phi_2(Y) := \mathbb{V}[Y], \dots$ . Crucially, we note that *our approximations can only relax the constraints*, i.e., the optima may result in looser bounds compared to the full constraint set, *but not invalid bounds*, barring bad local optima as well as Monte Carlo and estimation errors.

## 4 Optimization Strategy

Here we state our final non-convex, yet smooth, constrained optimization problem:

$$\begin{aligned} \text{objective:} & & o_{x^*}(\eta) & := \int f_\theta(x^*) p_\eta(\theta) d\theta \\ \text{constraint LHS:} & & \text{LHS}_{m,l} & := \mathbb{E}[\phi_l(Y) | z^{(m)}] \\ \text{constraint RHS:} & & \text{RHS}_{m,l}(\eta) & := \int \phi_l(f_\theta(x)) p_\eta(x, \theta | z^{(m)}) dx d\theta \\ \text{opt. problem:} & & \min_{\eta} / \max_{\eta} o_{x^*}(\eta) & \quad \text{s.t.} \quad \text{LHS}_{m,l} = \text{RHS}_{m,l}(\eta) \text{ for all } m \in [M], l \in [L] \end{aligned}$$

Here, min and max give the lower and upper bound respectively. In this section we describe how to tackle the optimization with an augmented Lagrangian strategy (Nocedal & Wright, 2006) and how to estimate all quantities from observed data. Algorithm 1 in Appendix D describes the full procedure.

<sup>6</sup>A full discussion on the construction and implications of such assumptions is given in Appendix B.

<sup>7</sup>We discuss discrete outcomes or discrete features, which could also lead to discontinuous  $f_\theta$  in Appendix C.

## 4.1 Augmented Lagrangian Strategy

We can think of the left-hand side LHS as target values, estimated once up front from observed data. The right-hand side RHS is estimated repeatedly using samples from our model  $p_\eta(x, \theta | z^{(m)})$  during optimization. For notational simplicity, we will often “flatten” the indices  $m$  and  $l$  into a single index  $l \in [M \cdot L]$ . Since LHS is subject to misspecification and estimation error, we introduce positive tolerance variables  $b \in \mathbb{R}_{>0}^{M \cdot L}$ , relaxing equality constraints into inequality constraints

$$c_l(\eta) := b_l - |\text{LHS}_l - \text{RHS}_l(\eta)| \geq 0, \quad \text{with } b_l := \max\{\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \cdot |\text{LHS}_l|\},$$

for fixed absolute and relative tolerances  $\epsilon_{\text{abs}}, \epsilon_{\text{rel}} > 0$ . The constraint  $c_l(\eta)$  is satisfied if  $\text{RHS}_l(\eta)$  is *either* within a fraction  $\epsilon_{\text{rel}}$  of  $\text{LHS}_l$  *or* within  $\epsilon_{\text{abs}}$  of  $\text{LHS}_l$  in absolute difference. The absolute tolerance is useful when LHS is close to zero. The exact constraints are recovered as  $\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \rightarrow 0$ . Again, the introduced tolerance can only make the obtained bounds looser, not invalid.

We consider an inequality-constrained version of the augmented Lagrangian approach with Lagrange multipliers  $\lambda \in \mathbb{R}^{M \cdot L}$  (detailed in Section 17.4 of Nocedal & Wright (2006)). Specifically, the Lagrangian we aim to minimize with respect to  $\eta$  is:

$$\mathcal{L}(\eta, \lambda, \tau) := \pm_{o_{x^*}}(\eta) + \sum_{l=1}^{M \cdot L} \begin{cases} -\lambda_l c_l(\eta) + \frac{\tau c_l(\eta)^2}{2} & \text{if } \tau c_l(\eta) \leq \lambda_l, \\ -\frac{\lambda_l^2}{2\tau} & \text{otherwise,} \end{cases} \quad (7)$$

where  $+/-$  is used for the lower/upper bound and  $\tau$  is a temperature parameter, which is increased throughout the optimization procedure. Given an approximate minimum  $\eta$  of this subproblem, we then update  $\lambda$  and  $\tau$  according to  $\lambda_l \leftarrow \max\{0, \lambda_l - \tau c_l(\eta)\}$  and  $\tau \leftarrow \alpha \cdot \tau$  for all  $l \in [M \cdot L]$  and a fixed  $\alpha > 1$ . The overall strategy is to iterate between minimizing eq. (7) and updating  $\lambda_l$  and  $\tau$ .

## 4.2 Empirical Estimation and Implementation Choices

For a dataset  $\mathcal{D} = \{(z_i, x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^3$ , we describe our method in Algorithm 1 in Appendix D.

**Pre-processing.** As a first step, we whiten the data (subtract mean, divide by variance). Then, we interpolate the CDF  $\hat{F}_Z$  of  $\{z_i\}_{i=1}^N$  to compute the grid points  $z^{(m)}$ . Next, we assign each observation to a grid point via  $\text{bin}(i) := \max\{\arg \min_{m \in [M]} |z_i - z^{(m)}|\}$  for  $i \in [N]$ , i.e., each datapoint is assigned to the gridpoint that is closest to its  $z$ -value (higher bin for ties). Given  $M, L$  and  $\phi_l$ , we can estimate  $\text{LHS}_{m,l}$  from data via  $\text{LHS}_{m,l} := \mathbb{E}[\phi_l(Y) | z^{(m)}] \approx \frac{1}{|\text{bin}^{-1}(m)|} \sum_{i \in \text{bin}^{-1}(m)} \phi_l(y_i)$ , which remain unchanged throughout the optimization. This allows us to fix the tolerances  $b = \max\{\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \text{LHS}\}$ . Finally, we obtain a single batch of examples from  $X | z^{(m)}$  of size  $B \in \mathbb{N}$ , which we will also reuse throughout the optimization via inverse CDF sampling  $\hat{x}_j^{(m)} = \hat{F}_{X|z^{(m)}}^{-1}(j^{-1}/B - 1)$  for  $j \in [B], m \in [M]$ . Here,  $\hat{F}_{X|z^{(m)}}$  is the CDF of  $\{x_i\}_{i \in \text{bin}^{-1}(m)}$ .






**Monte Carlo estimation.** To minimize the Lagrangian, we use stochastic gradient descent (SGD). Therefore, we need to compute (estimates for)  $\nabla_\eta o_{x^*}(\eta), \nabla_\eta c_l(\eta)$ , where the latter boils down to  $\nabla_\eta \text{RHS}_{m,l}(\eta)$ . In practice, we compute Monte Carlo estimates of  $o_{x^*}(\eta)$  and  $\text{RHS}_{m,l}(\eta)$  and use automatic differentiation, e.g., using JAX (Bradbury et al., 2018), to get the gradients. If we had a batch of independent samples  $\theta^{(j)} \sim p_\eta(\theta)$  of size  $B$ , we could estimate the objective eq. (3) for a given  $\eta$  via  $\mathbb{E}[Y | do(x^*)] \approx \frac{1}{B} \sum_{j=1}^B f_{\theta^{(j)}}(x^*)$ . Similarly, with i.i.d. samples  $\theta^{(j)} \sim p_\eta(\theta | z^{(m)})$  we can estimate  $\text{RHS}_{m,l}$  in eq. (6) as  $\text{RHS}_{m,l}(\eta) \approx \frac{1}{B} \sum_{j=1}^B \phi_l(f_{\theta^{(j)}}(\hat{x}_j^{(m)}))$ . Hence, the last missing piece is to sample from eq. (4) in a fashion that maintains differentiability w.r.t.  $\eta$ . We follow the standard procedure to sample from a Gaussian copula for the parameters  $\theta^{(j)}$ , with the additional restriction to preserve the pre-computed sample  $\hat{x}$ . Algorithm 2 in Appendix D describes the sampling process from  $p_\eta(\theta, X | z^{(m)})$  as defined in Section 3.2 in detail. The output is a  $(K + 1) \times B$ -matrix, where the first row contains  $B$  independent  $X$ -samples and the remaining  $K$  rows are the components of  $\theta \in \mathbb{R}^K$ . We pool samples from all  $z^{(m)}$  to obtain samples from  $p_\eta(\theta)$ . By change of variables, the parameters  $\eta = (\mu, \sigma^2, L)$  enter in a differentiable fashion (c.f. reparameterization trick (Kingma & Welling, 2014)). We initialize  $\eta$  randomly, described in detail in Appendix D.4.


**Response functions.** For the family of response functions we first consider polynomials, i.e.,  $\psi_k(x) = x^{k-1}$  for  $k \in [K]$ . We will specifically focus on linear ( $K = 2$ ), quadratic ( $K = 3$ ), and



cubic ( $K = 4$ ) response functions. Second, we consider *neural basis functions (MLP)*, where we fit a multi-layer perceptron with  $K$  neurons in the last hidden layer to the observed data  $\{(x_i, y_i)\}_{i \in N}$  and take  $\psi_k(x)$  to be the activation of the  $k$ -th neuron in the last hidden layer. We describe the details as well as an additional choice based on Gaussian process basis functions in Appendix E.





The choice of polynomials is mainly done here to illustrate a type of sensitivity analysis: we will contrast how bounds change when moving from a linear to quadratic, then quadratic to cubic and learned MLP representations. Recall that finite linear combinations of basis functions can arbitrarily approximate infinite-dimensional function spaces. The practitioner should be free to choose its complexity and pay its price by getting less informative bounds. For instance, we can add as many knot positions for a regression splines procedure as we want to get arbitrarily close to nonparametric function spaces. There is no concern for overfitting, given that data plays a role only via the estimation of  $p(x, y | z)$  or of particular black-box expectations (Appendix G). We emphasize that *having a class of algorithms that allows for controlling the complexity of the function space is an asset, not a liability*. Knowledge of functional constraints is useful even in a non-causal setting (Gupta et al., 2020). The linear basis formulation can be as flexible as needed, while allowing for shape and smoothness assumptions that are more expressive than all-or-nothing assumptions about, say, missing edges or additivity. In Appendix F, we discuss an alternative based on discretization of  $X$  combined with the off-the-shelf use of Balke & Pearl (1994). We demonstrate how in several ways that is just a *less* flexible family of response functions than the approach discussed here, although see Appendix B for a discussion on making  $p_\eta(\theta | x, z)$  also more flexible than the implementation discussed here.

## 5 Experimental Results

We evaluate our method on a variety of synthetic and real datasets. In all experiments, we report the results of two stage least squares (**2SLS** ) and kernel instrumental variable regression (**KIV** ) (Singh et al., 2019). Note that both methods assume additive noise and provide point estimates for expected outcomes under a given treatment. The KIV implementation by Singh et al. (2019) comes as an off-the-shelf method with internal heuristics for tuning hyperparameters. For our method, we show **lower** () and **upper** () bounds computed individually for multiple values of  $x^* \in \mathbb{R}$ . The transparency of these lines indicates the tolerances  $\epsilon_{\text{abs}}, \epsilon_{\text{rel}}$ , where more transparency corresponds to larger tolerances. Missing bounds at an  $x^*$  indicate that the constraints could not be satisfied in the optimization. In the synthetic settings, we also show the **true causal effect**  $\mathbb{E}[Y | do(X = x^*)]$  ()

Finally, we highlight that there are multiple possible causal effects compatible with the data (which our method aims to bound). To do so, we fit a latent variable model of the form shown in Figure 1(a) to the data, with  $U | Z, X, Y \sim \mathcal{N}(\mu(Z, X, Y), \sigma^2(Z, X, Y))$  where  $\mu, \sigma^2$  as well as  $\mathbb{E}[X | Z, U]$  are parameterized by neural networks. We ensure that the form of  $\mathbb{E}[Y | X, U]$  matches our assumptions on the function form of the response family (i.e., either polynomials of fixed degree in  $X$ , or neural networks). We then optimize the evidence lower bound following standard techniques (Kingma & Welling, 2014), see Appendix H. We fit multiple models with different random initializations and compute **the implied causal effect** of  $X$  on  $Y$  for each one, shown as multiple thin gray lines () . We report results for additional datasets as well as how our method performs in the small data regime in Appendix I. All experiments use a single set of hyperparameters, which we describe in Appendix I.1.

**Linear Gaussian case.** First, we test our method in a synthetic linear Gaussian scenario, where instrument, confounder, and noises  $Z, C, e_X, e_Y$  are independent standard Gaussian variables. We consider two settings of the form  $X = g(Z, C, e_X) := \alpha Z + \beta C + e_X$  and  $Y = f(X, C, e_Y) := X - 6C + e_Y$ , with  $\alpha, \beta \in \{(0.5, 3), (3, 0.5)\}$ . The two settings of coefficients  $\alpha, \beta$  describe a weak instrument with strong confounding and a strong instrument with weak confounding respectively. The first two rows of Figure 2 show our bounds in these settings for linear, quadratic and MLP response functions. Because these scenarios satisfy all theoretical assumptions of 2SLS and KIV, 2SLS () reliably recovers the true causal effect, which is simply  $\mathbb{E}[Y | do(X = x^*)] = x^*$ . For a weak instrument, KIV () fails by reverting to its prior mean 0 everywhere, whereas it matches the true effect in data rich regions in the second setting with weak confounding.<sup>8</sup>

We observe that the true causal effect () is always within our bounds (, ) . Moreover, our bounds also contain most of the “other possible models” that could explain the data () , showing

<sup>8</sup>We provide more details on this failure mode of KIV in Appendix J.

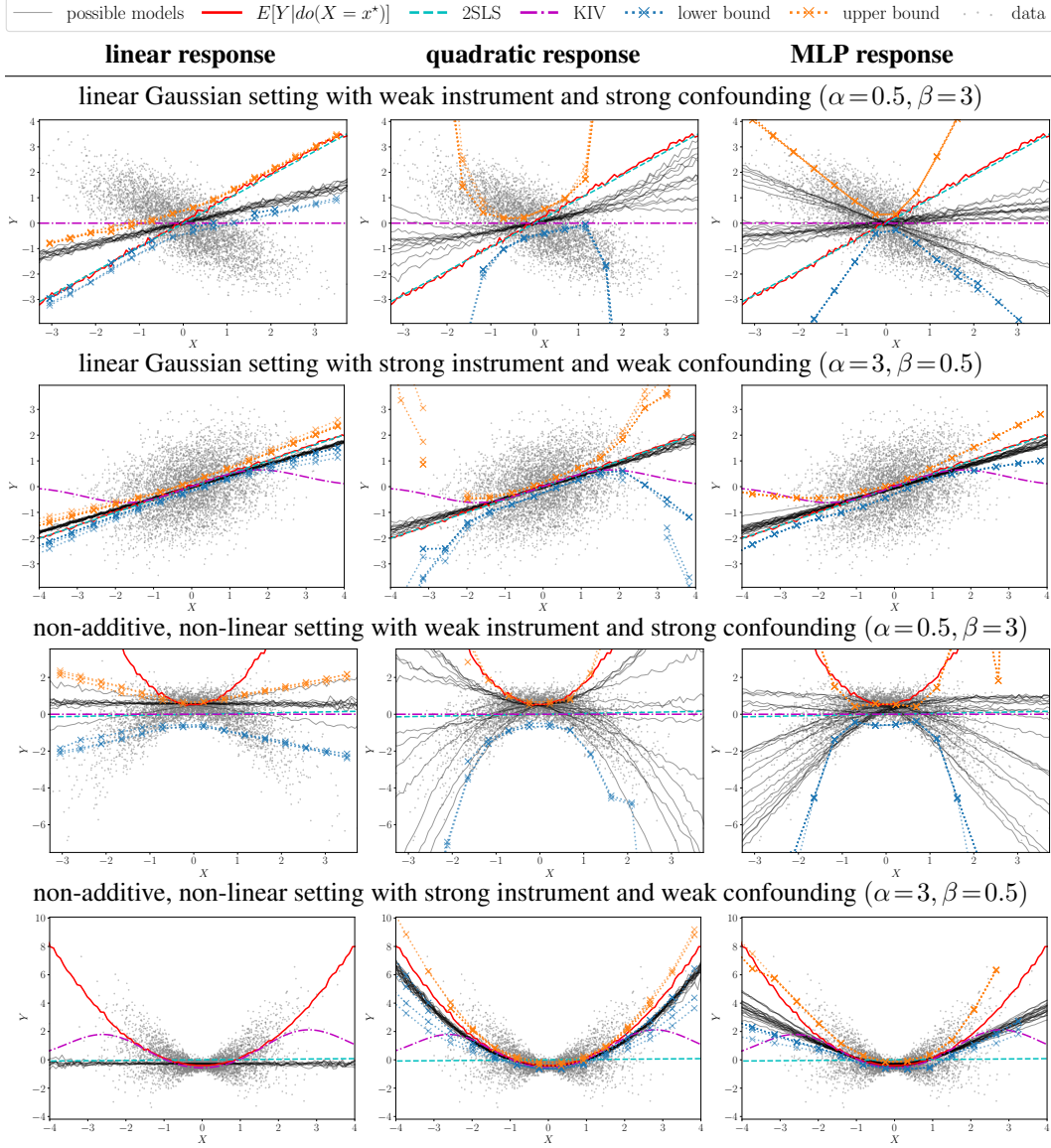


Figure 2: Results for synthetic datasets (linear Gaussian and non-linear, non-additive) for a weak and strong instrument respectively. Columns correspond to different response function families.

that they are highly informative, without being more confident than warranted. As expected, our bounds get looser as we increase the flexibility of the response functions (linear, quadratic, MLP from columns 1-3). In particular, allowing for flexible MLP responses (column 3), our bounds are rightfully loose for strong confounding. As confounding weakens and the instrument strengthens (in the second row) the gap between our bounds gets narrower.

**Non-additive, non-linear case.** Our next synthetic setting is non-linear and violates the additivity assumption. Again, the treatment is given by  $X = \alpha Z + \beta C + e_X$  with the same set of coefficients  $\alpha, \beta$  as for the linear setting. The outcome is non-linear and non-additive  $Y = 0.3 X^2 - 1.5 X C + e_Y$  with a true effect of  $\mathbb{E}[Y | do(X = x^*)] = 0.3 (x^*)^2$ . The bottom two rows of Figure 2 show our results for this setting. Since additivity is violated (due to the  $X C$ -term) and the effect is non-linear, 2SLS fails. Without additivity, KIV also fails for strong confounding, but captures the true effect well in data rich regions when the instrument is strong and confounding is weak. The strongly confounded case (row 3) highlights the effect of the choice of response functions. Wrongly assuming linear response functions, our bounds rule out the true effect (row 3, column 3). However, they capture the implied causal effects from possible compatible linear models. As we allow for more flexible



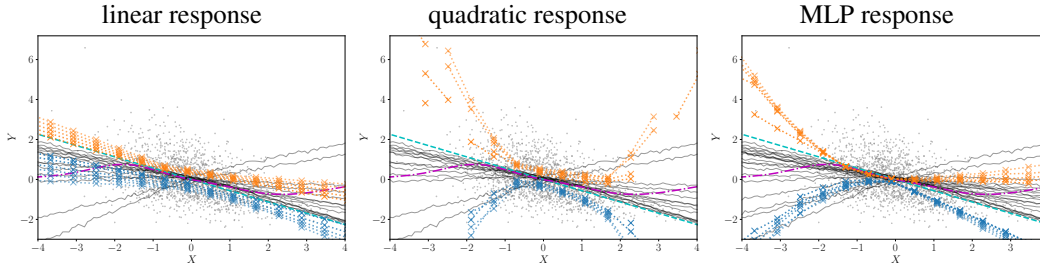


Figure 3: Results on the expenditure dataset for different response function families.

response functions capable of describing the true effect, our bounds are extremely conservative (row 3, columns 2 & 3) as they should be, indicated by the effects from other compatible models. In the strong instrument, weak confounding case (row 4), our bounds become narrower to the point of essentially identifying the true effect for adequate response functions (column 2). Here, linear response functions cannot explain the data anymore, indicated by missing bounds (row 4, column 1).

**Expenditure data.** We now turn to a real dataset from a 1995/96 survey on family expenditure in the UK (Office for National Statistics, 2000). This dataset has been used by Gunsilius (2020) and previously (Blundell et al., 2007; Imbens & Newey, 2009) for 1994/95 data. The outcome of interest is the share of expenditure on food. The treatment is the log of the total expenditure and the instrument is gross earnings of the head of the household. All three variables are continuous, relations cannot be expected to be linear, and we cannot exclude unobserved confounding (Gunsilius, 2020), making this a good test case for our method. We describe the data in more detail in Appendix I.3. Figure 3 shows that our bounds provide useful information about both the sign and magnitude of the causal effect and gracefully capture the increasing uncertainty as we allow for more flexible response functions. Moreover, they include most of the possible effects from latent variable models indicating that they are not overly restrictive. The few curves that escaped our bounds correspond to situations where the latent variable model fit was suboptimal in terms of local likelihood and hence may be an artifact of the latent variable model training procedure.

## 6 Conclusion

We have proposed a class of algorithms for computing bounds on causal effects by exploiting modern optimization machinery. While this addresses an important source of uncertainty in causal inference — partial identifiability as opposed to full identifiability — there is also statistical uncertainty: confidence or credible intervals for the *bounds* themselves (Imbens & Manski, 2004). Clearly this is an important matter to be addressed in future work, and the black-box approach of Silva & Evans (2016) provides some directions for credible intervals. There are also considerations about the parameterization of  $p_{\eta}(\theta | x, z)$  and how possible pre-treatment covariates can be non-trivially used in the model. We defer these considerations to Appendix B. Alternative parameterizations of the IV model, such as the one by Zhang & Bareinboim (2020) can lead to alternative algorithms and ways of expressing assumptions.

One could also use the same ideas to test whether an IV model is valid, one of the original motivations for deriving the implied constraints of latent variable causal models (e.g., Wolfe et al., 2019). In all that followed, we assumed that the model was correct. Model falsification can still be done, which will happen when the optimization fails to find a solution (Silva & Evans, 2016), and observed in some of the experiments reported. Further formalizing and specializing methods for testing models instead of deriving bounds is an interesting direction for future work.

Finally, we foresee our ideas as ways of liberating causal modeling to accommodate “softer,” more general constraints than conditional independence statements. For instance, as described by Silva & Evans (2016), there is no need to assume any sparsity in a causal DAG, as long as we know that some edges are “weak” (in a technical sense) so that, e.g., edge  $Z \rightarrow Y$  is allowed, but its influence on  $Y$  is not arbitrary. How to do that in a computationally feasible way remains a challenge, but the possibility of complementing causal inference based on sparse DAGs, such as the do-calculus of Pearl (2009), with the sledgehammer of modern continuous optimization, is an attractive prospect.

## Broader Impact

Cause effect estimation is crucial in many areas where data-driven decisions may be desirable such as healthcare, governance or economics. These settings commonly share the characteristic that experimentation with randomized actions is unethical, infeasible or simply impossible. One of the promises of causal inference is to provide useful insights into the consequences of hypothetical actions based on observational data. However, causal inference is inherently based on assumptions, which are often untestable. Even a slight violation of the assumptions may lead to drastically different conclusions, potentially changing the desired course of action. Especially in high-stakes scenarios, it is thus indispensable to thoroughly challenge these assumptions.

This work offers a technique to formalize such a challenge of standard assumptions in continuous IV models. It can thus help inform highly-influential decisions. One important characteristic of our method is that while it can provide informative bounds under certain assumptions on the functional form of effects, the bounds will widen as less prior information supporting such assumptions is available. We can view this as a way of deferring judgment until stricter assumptions have been assessed and verified.

Since our algorithms are causal inference methods, they requires assumptions too. Therefore, our method also requires a careful assessment of these assumptions by domain-experts and practitioners. In addition, as we are optimizing a non-convex problem with local methods, we have no theoretical guarantee of correctness of our bounds. Hence, if wrong assumptions for our model are accepted prematurely, or our optimization strategy fails to find global optima, our method may wrongly inform decisions. If these are high-stakes decisions, then wrong decisions can have significant negative consequences (e.g., a decision not to treat a patient that should be treated). If the data that this model is trained on is biased against certain groups (e.g., different sexes, races, genders) this model will replicate those biases. We believe a fruitful approach towards making our model more sensitive to uncertainties due to structurally-biased, unrepresentative data, is to learn how to derive, then inflate (to account for bias) uncertainty estimates for our bounds.

## Acknowledgments and Disclosure of Funding

We thank Florian Gunsilius for useful discussions, providing code for his method and explaining how to prepare the Family Expenditure Survey dataset. We are grateful to Robin Evans, Arthur Gretton, Jiri Hron, Paul Rubenstein, and Rahul Singh for useful discussions and feedback. MK and RS acknowledge support from the The Alan Turing Institute under EPSRC grant EP/N510129/1. This work was partially done while RS was on a sabbatical at the Department of Statistics, University of Oxford.

## References

- Acemoglu, D., Johnson, S., and Robinson, J. A. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401, 2001.
- Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press, 2008.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.
- Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pp. 46–54. Morgan Kaufmann Publishers Inc., 1994.
- Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pp. 3564–3574, 2019.
- Blundell, R., Chen, X., and Kristensen, D. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

- Bonet, B. Instrumentality tests revisited. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pp. 48–55, 2001.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- Chen, X. and Christensen, T. M. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.
- Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.
- Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- Drton, M., Sturmfels, B., and Sullivant, S. *Lectures on Algebraic Statistics (Oberwolfach Seminars Book 39)*. Springer, 2009.
- Evans, R. Margins of discrete Bayesian networks. *Annals of Statistics*, 46(6A):2623–2656, 2019.
- Gunsilius, F. Testability of instrument validity under continuous endogenous variables. *arXiv preprint arXiv:1806.09517*, 2018.
- Gunsilius, F. A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv preprint arXiv:1910.09502*, 2020.
- Gupta, M. R., Loudior, E., Mangylov, O., Morioka, N., Narayan, T., and Zhao, S. Multidimensional shape constraints. *Proceedings of the Thirty-Seven International Conference in Machine Learning (ICML2020)*, 2020.
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1414–1423. JMLR. org, 2017.
- Hestenes, M. R. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.
- Horowitz, J. L. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2): 347–394, 2011.
- Imbens, G. and Manski, C. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.
- Imbens, G. W. and Newey, W. K. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.
- Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. The sensitivity of counterfactual fairness to unmeasured confounding. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 213. AUAI Press, 2019.
- Kingma, P. D. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1, 2014.
- Lewis, G. and Syrgkanis, V. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.
- Manski, C. *Identification for Prediction and Decision*. Harvard University Press, 2007.
- Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 2020.
- Navascues, M. and Wolfe, E. The inflation technique solves completely the classical inference problem. *arXiv: 1707.06476*, 2019.

- Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.
- Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.
- Office for National Statistics. Family expenditure survey, 1996-1997, 2000. URL <http://doi.org/10.5255/UKDA-SN-3783-1>.
- Palmer, T., Ramsahai, R., Didelez, V., and Sheehan, N. Nonparametric bounds for the causal effect in a binary instrumental variable model. *The Stata Journal*, 11(3):345–367, 2011.
- Pearl, J. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443. Morgan Kaufmann Publishers Inc., 1995.
- Pearl, J. *Causality*. Cambridge university press, 2009.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Silva, R. and Evans, R. Causal inference through a witness protection program. *Journal of Machine Learning Research*, 17:1–53, 2016.
- Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pp. 4595–4607, 2019.
- Tian, J. and Pearl, J. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 519–527, 2002.
- VanderWeele, T. and Hernán, M. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.
- Wolfe, E., Spekkens, R. W., and Fritz, T. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.
- Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT press, 2010.
- Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcomes. *Columbia CausalAI Laboratory, Technical Report (R-61)*, 2020.

## A Gunsilius’s Algorithm

Gunsilius (2020) provides a theoretical framework for minimal conditions for a continuous IV model to imply non-trivial bounds (that is, bounds tighter than what can be obtained by just assuming that the density function  $p(x, y | z)$  exists). That work also introduces two variations of an algorithm for fitting bounds.

The basic version consists of first sampling  $l$  response functions  $f_{R_x}(\cdot)$  and  $f_{R_y}(\cdot)$  from a distribution over functions – in the experiments described, a Gaussian process evaluated on a grid in the respective spaces. The final distribution is a reweighted combination of the pre-sampled  $l$  response functions with weights  $\mu$  playing the role of the decision variables to be optimized. Hence, by construction, the space of distributions in the response function space is absolutely continuous with respect to the pre-defined Gaussian process. The constraints are defined by approximating an estimate of the bivariate CDF  $F(x, y | z)$  on a grid of values, which are approximately constrained to match the model implied CDF in a  $L_2$  sense. Large deviance bounds are then used to show the (intuitive) result that this approximation is a probably approximately correct formulation of the original optimization problem.

One issue with this algorithm is that  $l$  may be required to be large as it is a non-adaptive Monte Carlo approximation in a high dimensional space. A variant is described where, every time a solution for  $\mu$  is found, response function samples with low corresponding values of  $\mu$  are replaced (again, from the given and non-adaptive Gaussian process). Although this now has the advantage of adapting the Monte Carlo samples to the problem, this has convergence problems that may be severe and not easy to diagnose.

In contrast, we formulate our adaptation of  $\eta$  as a continuous optimization problem with an estimate of the gradient that has empirically reasonable stability, as expected from the related work in the machine learning literature for gradient estimation. We also parameterize the distribution so that the only constraint that we need to enforce concerns the univariate density  $p(y | z)$  (or  $p(y | x, z)$ , in the variation discussed in Appendix G, which in principle requires no density estimation). Like the algorithm given by Gunsilius, the space of functions is a linear combination of a fixed dictionary of basis functions with a Gaussian distribution on the parameters, although we do not make use of the discrete mixture reweighting on the Monte Carlo samples, which introduces instability in (Gunsilius, 2020) despite its good theoretical properties. Our formulation, like the one in (Gunsilius, 2020), can in principle make use of more flexible distributions such as a mixture of Gaussian copulas at the cost of more computation, as discussed in Appendix B. An important piece of future work is to thoroughly assess how stable a mixture of Gaussians version of our algorithm is in practice.

The proposed implementation of Gunsilius’ algorithm computes  $F_{Y | do(x_0^*)}(y^*) - F_{Y | do(x_1^*)}(y^*)$ , i.e., the difference in effects at two different treatment levels  $x_0^*$  and  $x_1^*$  for individuals within a fixed quantile  $y^* \in [0, 1]$  of the outcome variable. For example, in the expenditure dataset (see Section I.3), the setting  $x_0^* = 0.75$ ,  $x_1^* = 0.25$ ,  $y^* = 0.25$  would look at how much people, who spend a lot overall ( $x^* = 0.75$ ) and spend comparably little on food (up to 25%), would spend on food relatively to overall expenditure, if they spent much less overall ( $x_1^* = 0.25$ ). The main tuning parameter in the proposed algorithm is the penalization parameter  $\lambda$ , which corresponds to the tightness of the constraint. In the proposed implementation, this parameter is fixed throughout the optimization and must be chosen manually. In Figure 4, we show the results of Gunsilius’s algorithm for three different levels of  $y^*$  on the expenditure dataset. Small values of  $\lambda$  result in uninformatively loose bounds and do not always seem to converge (e.g., for  $y^* = 0.75$ ). As we increase  $\lambda$ , which corresponds to stronger enforcement of the constraint, the bounds get narrower. However, even after a long burn-in period, we still encounter substantial “instantaneous jumps” as well as longer-term drifts in the bounds, which may change the qualitative conclusions (for example in the  $y^* = 0.75$  setting). Note that this algorithm works on the empirical CDFs of all variables, i.e., they are all scaled to lie within  $[0, 1]$ .

Moreover, even after laboriously improving the performance of the algorithm using acceleration via JAX (Bradbury et al., 2018) and parallelized solving of the quadratic programs with CVXPY (Diamond & Boyd, 2016), producing an upper and lower bound for a single setting of  $x_0^*$ ,  $x_1^*$ ,  $y^*$ ,  $\lambda$  with Gunsilius’s algorithm took longer (about 30 minutes on a quad-core Intel Core i7) than a full set of upper and lower bounds at 15 different  $x^*$  values with our algorithm (about 20 minutes on the same hardware).

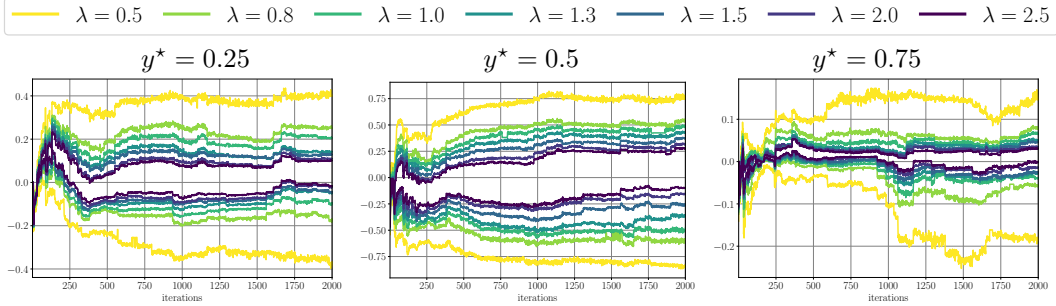


Figure 4: We show results of Gunsilius’s algorithm for 3 different settings of  $y^* \in \{0.25, 0.5, 0.75\}$ .

## B The Shape of $p_\eta(\theta | x, z)$ and Conditional Effects

It is not difficult to show that our parameterization of  $p_\eta(\theta | x, z)$  enforces  $\theta \perp\!\!\!\perp Z$  while allowing for  $\theta \not\perp\!\!\!\perp Z | X$ , as suggested by Figure 1(c). It follows directly by factoring a conditional density in terms of a copula density  $c(\cdot)$  and the required univariate marginals. That is, for some  $(V_1, V_2, V_3)$  for which we want to define a conditional pdf  $p(v_2 | v_1, v_3)$ , we have

$$\begin{aligned} p(v_1, v_2 | v_3) &:= c(F(v_1 | v_3), F(v_2 | v_3)) p(v_1 | v_3) p(v_2 | v_3) \quad \Rightarrow \\ p(v_2 | v_1, v_3) &= c(F(v_1 | v_3), F(v_2 | v_3)) p(v_2 | v_3). \end{aligned}$$

Since  $\int p(v_1, v_2 | v_3) dv_1 = p(v_2 | v_3)$ , a necessary and sufficient condition for  $V_2 \perp\!\!\!\perp V_3$  is choosing a model marginal such that  $p(v_2 | v_3) = p(v_2)$ . If  $c(F(v_1 | v_3), F(v_2))$  cannot be factored in terms of some product  $h_1(v_1, v_3)h_2(v_1, v_2)$ , which is typically the case, then  $V_2 \not\perp\!\!\!\perp V_3 | V_1$ .

The main apparent limitation of our  $p_\eta(\theta_k)$  (and the related copula) is its reliance on a parametric form. There is a complex relationship between the shape of the response function space and the distribution implied on that space by the unknown model  $\mathcal{M}$ . For  $Y = f(X, U)$ , it is always possible to assume without loss of generality that  $U$  is a set of variables which are marginally standard Gaussians: just let the transformation  $U_i' := \Phi^{-1}(F_i(U_i))$  be absorbed into  $f(\cdot)$ , where  $F_i(\cdot)$  is the marginal CDF of  $U_i$  and  $\Phi(\cdot)$  is the CDF of a standard Gaussian. Moreover, assuming that any dependence among elements of  $U$  can be explained by direct causation among them or by other latent parents, we can also assume all members of  $U$  are independent.

However, we do not want to assume a one-to-one correspondence between elements of  $\theta$  and elements of  $U$ : that is the whole point of using response functions. Even independent standard Gaussian  $U$ s would not translate to marginally Gaussian  $\theta$ . As an example, suppose  $Y = U_1^2 X + \lambda U_2$ . All response functions can be written in the form  $f_\theta(x) := \theta_1 x + \theta_2$ , where  $\theta_1 = U_1^2$  and  $\theta_2 = \lambda U_2$ . Hence,  $\theta_1$  follows a chi-squared distribution and  $\theta_2$  a zero-mean, but not standard, Gaussian. If  $Y = U_1 X^2 + \lambda U_1 U_2$ , then on top of that  $\theta_1$  and  $\theta_2$  are not independent.

The solution is conceptually not complicated: just let  $p_\eta(\cdot)$  be as flexible as desired. For instance, *let the copula be a finite or Dirichlet process mixture of Gaussian copulas, also defining flexible models for the marginals*. The IV conditional independence structure among  $Z, X, \theta$  is still preserved. The practical issue of course is the optimization. The algorithm of Gunsilius (2020) itself tries to approach the problem by learning the reweighting of a Monte Carlo approximation to a fixed base measure. That alone is already very computationally demanding and has convergence problems.

We set a parametric form for  $p_\eta(\cdot)$  for reasons beyond a compromise between flexibility and computational tractability. *Adopting a nonparametric model for the causal model, such as a Dirichlet process, seems pointless because:* i. we do not perform statistical inference directly in the causal model, but only via black-box estimators of (features of)  $p(x, y | z)$ , which can be nonparametric; ii. if we were to follow the route of performing statistical inference by directly fitting the causal model, the corresponding estimator would have a finite representation with dimensionality given by the data. A practical resource, sample size, limits the representational size of the estimator. The role of nonparametrics is to provide a type of adaptive regularization, and to provide theory about limits of parametric estimators as done by Gunsilius (2020). The latter has clear value in itself but it does not demand nonparametric models to be actually implemented, while the former is out of our

scope: in our case, no regularization is needed for the causal model as we do not fit data based on it. Instead, our practical resource is the computational budget: if we want to not use domain knowledge to perform the causal analysis, we simply choose the size of  $\eta$  based directly on the main bottleneck, the amount of computation available. Hence, by the time-data bounded nature of computational and statistical inference, we lose nothing by adopting a finite representation for both  $\eta$  and  $\theta$ .

The practitioner should be invited to sample from the implied function space to visualize whether the distribution of sample paths has a desired level of variability. Getting the “exact” shape of the true distribution is however nowhere as important as just having enough variability to avoid overconfident bounds. How to achieve “enough variability” without aiming at a completely flexible distribution of  $\theta$  may be a compromise between computational costs and domain-dependent judgment. However, in principle, the finite mixture of Gaussians approach can be done with the reparameterization trick. The relation to Gungilius algorithm is that our “base measure” is smoothly adaptive, leading to possibly more stable behaviour in practice. The price to be paid is that each iteration in our method would be substantially more expensive than the efficient mixture component weighting optimization done at each iteration of Gungilius’ method, *if* we were to optimize the mixture component parameters to completion while fixing the samples. However, we do joint partial optimization by gradient-informed small steps, taken at each sampling stage. This is one of the main distinctive features of our class of algorithms compared to the resample/optimize alternating procedure of Gungilius (2020).

To summarize, *the Gaussian case, discussed in the main text, should be seen as a useful illustration, not as a one-size-fits-all solution.* Any copula for which the reparameterization trick can be used can be automatically plugged into any instance of our class of algorithms.

Another important aspect brought by a parameterization of  $p_\eta(\cdot)$  is in case we have pre-treatment covariates  $W$  to either reduce confounding, remove (direct) dependence between  $Z$  and  $U$  or  $Z$  and  $Y$ , or just to answer questions related to conditional expected outcomes e.g.  $\mathbb{E}[Y | do(x), w]$  and conditional average causal effects (CATE),  $\mathbb{E}[Y | do(x), w] - \mathbb{E}[Y | do(x'), w]$ . Although a response function can straightforwardly depend on a vector of treatment variables, this makes less sense if variables  $W$  are not direct causes of  $Y$ . And even if elements of  $W$  are direct causes, we may want to treat them analogously to  $U$ : playing a role in the response function only via the distribution of  $\theta$ , instead of being explicitly in the scope of such functions.

*Modeling CATE can then be done in a completely straightforward way.* Nothing in the algorithm changes if we use a probabilistic model for  $p(x, y | z, w)$  to provide the observable counterpart of the causal model. Each configuration  $w$  defines a separate optimization problem. The corresponding factor  $p(\theta | x, z, w)$  can be set independently for each instance of  $w$ , regardless of its dimensionality.

However, a practitioner may be interested on providing information about how  $p(\theta | x, z, w)$  varies smoothly across values of  $w$  in order to impose further constraints on the response functions across multiple  $w$  realizations. We suggest that a way of incorporating covariates  $W$  is by a multilevel approach: define  $p_{\eta(w)}(\theta | x, z, w)$ , where each element of  $\eta$  may itself be a function of  $W$ , e.g.  $\mu_1 = \beta_1^\top W$  for some parameter vector  $\beta_1$ . Here,  $p(x | z, w)$  and  $p(y | z, w)$  (or  $p(y | x, z, w)$ ) are the marginals to be matched. We will discuss in future work ways of making  $p_\eta(\cdot)$  more flexible in general, including the use of covariates.

## C Discrete Outcomes and Discrete Features

If  $Y$  is discrete,  $f_\theta(x)$  will be discontinuous. Theoretically this will not pose a problem as long as the number of discontinuities is finite (Gungilius, 2020). The main practical issue is optimization, as eq. (6) will now not lead itself to gradient-based methods. The most immediate approximation is to use differentiable surrogates of  $f_\theta(x)$  that relax the constraints. In the most basic formulation, we have the inequalities

$$tol_- \leq \mathbb{E}[\phi_l(Y) | z^{(m)}] - \int \phi_l(f_\theta(x)) p_\eta(x, \theta | z^{(m)}) dx d\theta \leq tol_+,$$

for some tolerance factors  $tol_+, tol_-$ . Given upper and lower bounds  $\phi_l^+(f_\theta(x)), \phi_l^-(f_\theta(x))$  on  $\phi_l(f_\theta(x))$ , the relaxed constraints

$$\begin{aligned} tol_- &\leq \mathbb{E}[\phi_l(Y) | z^{(m)}] - \int \phi_l^-(f_\theta(x)) p_\eta(x, \theta | z^{(m)}) dx d\theta \\ \mathbb{E}[\phi_l(Y) | z^{(m)}] - \int \phi_l^+(f_\theta(x)) p_\eta(x, \theta | z^{(m)}) dx d\theta &\leq tol_+, \end{aligned}$$

will still result in valid, but looser bounds (again, up to local optima and Monte Carlo error). If  $f_\theta(x)$  is non-negative (for instance, if its codomain is  $\{0, 1\}$ ) and  $\phi_l(\cdot)$  is monotonic for non-negative inputs (such as  $\phi_l(x) = x$  and  $\phi_l(x) = x^2$ ), it is enough to plug in bounds for  $f_\theta(x)$  itself. We will elaborate on that in future work. In this context, we can also formulate an alternative approach to matching  $p(y | z)$ .

**Alternative Approach to Matching  $p(y | z)$ .** Here we describe an alternative approximation of eq. (5) that hinges on smoothly approximating the indicator function to render the integral well behaved. First, instead of evaluating  $\Pr(Y < y | Z = z^{(m)})$  for all  $y \in \mathcal{Y}$ , we take a similar approach for discretizing  $Y | z^{(i)}$  as we took for  $z^{(m)}$ . For a given  $z^{(m)}$ , instead of all half-spaces  $Y < y$ , we only consider the sets

$$A^{(m,l)} := (-\infty, y^{(m,l)}] \quad \text{with} \quad y^{(m,l)} := F_{Y|z^{(m)}}^{-1}\left(\frac{l-1}{L-1}\right)$$

for  $l \in [L]$  with some fixed  $L \in \mathbb{N}$ . This results in constraints for the  $L$ -quantiles of the conditional distributions of  $Y$

$$\frac{l-1}{L-1} = \int \mathbf{1}\left(f_\theta(x) \leq y^{(m,l)}\right) p_\eta(x, \theta | z^{(m)}) dx d\theta.$$

for all  $m \in [M]$  and  $l \in [L]$ . In practice, we would evaluate the integral on the right hand side with a Monte Carlo estimate, sampling from  $p_\eta(x, \theta | z^{(m)})$  and then differentiate with respect to  $\eta$  for gradient-based optimization. Therefore, the non-differentiable (even non-continuous) indicator function poses an issue for the optimization. We can circumvent this problem by approximating the indicator with a smoothly differentiable function, for example

$$\mathbf{1}(t \leq t^*) \approx \sigma_\rho(t - t^*) \quad \text{for} \quad \sigma_\rho(t) := \frac{1}{1 + e^{-\rho t}} \quad \text{or} \quad \sigma_\rho(t) := \frac{1}{1 + \exp\left(-\rho\left(t + \frac{1}{\sqrt{\rho}}\right)\right)}$$

for  $\rho > 0$ . As  $\rho \rightarrow \infty$ ,  $\sigma_\rho(t) \rightarrow \mathbf{1}(t \leq 0)$  pointwise on  $\mathbb{R} \setminus \{0\}$ , i.e., we can slowly increase  $\rho$  throughout the optimization to gradually approximate the constraints.

Hence an alternative approach to implement the constraint for matching  $p(y | z)$  is

$$\frac{l-1}{L-1} = \int \sigma_\rho(f_\theta(x) - y^{(m,l)}) p_\eta(x, \theta | z^{(m)}) dx d\theta$$

for all  $m \in [M]$  and  $l \in [L]$ , where we increase  $\rho > 0$  after each optimization round.

In practice, we this approach gave less robust results than the approach described in the main text, partly due to the additional hyperparameter schedule needed for  $\rho$ . Therefore, we only report results for the approach using dictionary functions  $\phi_l$  described in the main text.

## D Algorithm

### D.1 Additional Details of the Optimization

**Smoothen LHS.** Since  $\text{LHS}_{m,l}$  are estimated via empirical averages of  $\phi_l(y_i)$  for datapoints in a given bin  $i \in \text{bin}^{-1}(m)$ , “neighbouring” constraints  $\text{LHS}_{m,l}$  and  $\text{LHS}_{m+1,l}$  may have substantially different values. Since our model is smooth, it can be hard to match such non-continuities with  $\text{RHS}_{m,l}(\eta)$ . Intuitively, we expect such jumps to be artifacts of finite sample effects and not important properties of the true data distribution. Hence we apply a spline regression to the values  $\{\text{LHS}_{m,l}\}_{m=1}^M$  for each  $l \in [L]$  to smoothen out larger jumps between neighbouring values. In practice, we use a cubic univariate spline for each  $l$  with a smoothing factor of 0.2.

### D.2 Augmented Lagrangian Optimization Strategy

The Augmented Lagrangian method (Hestenes, 1969) is a general method for constrained optimization, originally proposed just for dealing with equality constraints. The benefit of this over penalty methods is that we do not need to take the penalty parameters  $\tau$  to  $\infty$  in order to solve the original constrained optimization problem, which can cause ill-conditioning (Nocedal & Wright, 2006).



---

**Algorithm 1** Bounding the IV interventional effect at treatment level  $x^*$ .

**Require:** dataset  $\mathcal{D} = \{(z_i, x_i, y_i)\}_{i=1}^N$ ; number of  $z$  grid points  $M$ ; constraint functions  $\{\phi_l\}_{l=1}^L$ ; response function family  $\{f_\theta\}_{\theta \in \Theta}$ ; batchsize  $B$ ; initial temperature  $\tau^{(0)} > 0$ ; temperature increase factor  $\alpha > 1$ ; tolerances  $\epsilon_{\text{abs}}, \epsilon_{\text{rel}}$ ; initial Lagrange multipliers  $\lambda$ ; initial parameters  $\eta^{(0)}$ ;

- 1:  $z^{(m)} := \hat{F}_Z^{-1}(\frac{m}{M+1})$  for  $m \in [M]$  ▷  $\hat{F}_Z$ : CDF of  $\{z_i\}_{i=1}^N$ .
- 2:  $\text{bin}(i) := \max\{\arg \min_{m \in [M]} |z_i - z^{(m)}|\}$  for  $i \in [N]$  ▷ split data points into “z-bins”
- 3:  $\text{LHS}_{m,l} := \frac{1}{|\text{bin}^{-1}(m)|} \sum_{i \in \text{bin}^{-1}(m)} \phi_l(y_i)$  for  $m \in [M], l \in [L]$  ▷ pre-compute LHS
- 4: smoothen  $\text{LHS}_{m,l}$  across  $m$  for each  $l$  with spline regression ▷ see Appendix D.1
- 5:  $b := \max\{\epsilon_{\text{abs}}, \epsilon_{\text{rel}} \text{LHS}\}$  (element-wise) ▷ set constraint tolerances
- 6:  $\hat{x}_j^{(m)} := \hat{F}_{X|z^{(m)}}^{-1}(\frac{j-1}{B-1})$  for all  $j \in [B], m \in [M]$  ▷  $\hat{F}_{X|z^{(m)}}$ : CDF of  $\{x_i\}_{i \in \text{bin}^{-1}(m)}$
- 7: **for**  $t = 1 \dots T$  (or until convergence) **do** ▷ optimization rounds
- 8:    $\eta^{(t)} := \text{OPTIMIZE SUBPROBLEM}(\eta^{(t-1)}, \lambda^{(t-1)}, \tau^{(t-1)})$  ▷ min. Lagrangian at fixed  $\lambda, \tau$
- 9:    $\lambda_l^{(t)} \leftarrow \max\left(0, \lambda_l^{(t-1)} - \tau^{(t-1)} c_l(\eta^{(t)})\right)$  ▷ update Lagrangian multipliers
- 10:    $\tau^{(t)} \leftarrow \alpha \tau^{(t-1)}$  ▷ increase temperature parameter
- 11: **return**  $o_{x^*}(\eta^{(T)})$

12: **function**  $\text{OPTIMIZE SUBPROBLEM}(\eta, \lambda, \tau)$

- 13:   ▷ In here we use SGD with auto-differentiation to minimize  $\mathcal{L}$ . Hence we only describe how to evaluate  $\mathcal{L}$  in a differentiable fashion:
- 14:    $o_{x^*}(\eta) := \frac{1}{B} \sum_{j=1}^B f_{\theta^{(j)}}(x^*)$  with  $\theta^{(j)} \sim p_\eta(\theta)$  ▷ c.f. Algorithm 2 for sampling
- 15:    $\text{RHS}_{m,l}(\eta) := \frac{1}{B} \sum_{j=1}^B \phi_l(f_{\theta^{(j)}}(\hat{x}_j^{(m)}))$  ▷ c.f. Algorithm 2 for sampling
- 16:    $c(\eta) := b - |\text{LHS} - \text{RHS}(\eta)|$  ▷ compute constraint terms
- 17:    $\mathcal{L}(\eta) := \pm o_{x^*}(\eta) + \sum_{l=1}^{M \cdot L} \xi(c_l(\eta), \lambda_l, \tau)$  ▷ Lagrangian ( $\pm$  for lower/upper bound)
- 18:   **return**  $\arg \min_\eta \mathcal{L}(\eta)$  ▷ optimize with SGD

---

However, our problem only contains inequality constraints. Thus, we consider a refinement proposed by Nocedal & Wright (2006) to purely handle inequality constraints using Augmented Lagrangian methods. Specifically, we can write the inequality constrained optimization problem equivalently as an unconstrained optimization problem with Lagrange multipliers  $\lambda$ :

$$\min_{\eta} \max_{\lambda \geq 0} \left\{ o_{x^*}(\eta) + \lambda^\top (c(\eta) - b) \right\}.$$

To see that it is equivalent, note that the max returns  $o(\eta)$  when  $\eta$  satisfies the constraints (as the maximum is obtained at  $\lambda = 0$ ), and  $\infty$  otherwise (as the maximum is at  $\lambda = \infty$ ). However, this is not easy to optimize as the  $\lambda$  jumps from 0 to  $\infty$  when passing through the constraint boundary. To fix this, we add a term that penalizes  $\lambda$  making larger changes from its previous value. Specifically,

$$\min_{\eta} \max_{\lambda \geq 0} \left\{ o(\eta) + \lambda^\top (c(\eta) - b) - \frac{1}{2\tau} \|\lambda - \lambda'\|^2 \right\},$$

where  $\lambda'$  are the Lagrange multipliers from the previous iteration and  $\tau$  is a penalty term that is iteratively increased. Note that the max optimization can be solved in closed form for each Lagrange multiplier  $\lambda_l$

$$\lambda_l = \max\{0, \lambda_l' + \tau c_l(\eta)\},$$

where  $c_l(\eta)$  is shorthand for the  $l$ -th inequality constraint. Plugging these values into the optimization problem, we arrive at

$$\min_{\eta} \mathcal{L}(\eta, \lambda, \tau) := o_{x^*}(\eta) + \sum_{l=1}^{M \cdot L} \xi(c_l(\eta), \lambda_l, \tau)$$

---

**Algorithm 2** Sampling parameter values  $\theta$  from  $p_\eta(\theta, X | z^{(m)})$ .

---

- 1: Sample each component of  $w \in \mathbb{R}^{K \times B}$  i.i.d. from a standard Gaussian.
  - 2: Prepend the vector  $(0, 1/B, \dots, 1)$  as the first row of  $w$ , resulting in  $w \in \mathbb{R}^{(K+1) \times N}$ .
  - 3: Allow for dependencies between components by multiplying with the Cholesky factor  $w \leftarrow L w$ .
  - 4: Normalize all values by applying the standard Gaussian CDF component wise,  $w \leftarrow \varphi_{0,1}(w)$ .
  - 5: Fix the marginals of  $\theta_k^{(j)}$  by applying the inverse CDF of a  $(\mu_k, \sigma_k^2)$ -Gaussian:  $\theta^{(j)} \leftarrow \varphi_{\mu_k, \sigma_k^2}^{-1}(w_{j+1})$  for  $j \in [K]$ . Here,  $w_{j+1}$  denotes the  $j + 1$ -st row of  $w$ .
  - 6: Sampling  $X$  via  $\hat{F}_X^{-1}(w_1)$  by design simply gives the pre-computed  $\hat{x}$ .
- 

with

$$\xi(c_l(\eta), \lambda_l, \tau) := \begin{cases} -\lambda_l c_l(\eta) + \frac{\tau c_l(\eta)^2}{2} & \text{if } \tau c_l(\eta) \leq \lambda_l, \\ -\frac{\lambda_l^2}{2\tau} & \text{otherwise,} \end{cases}$$

where  $\tau$  increases throughout the optimization procedure. Given an approximate solution  $\eta$  of this subproblem, we then update  $\lambda$  according to

$$\lambda_l \leftarrow \max\{0, \lambda_l - \tau c_l(\eta)\}$$

for all  $l \in [M \cdot L]$  and set  $\tau \leftarrow \alpha \tau$  for a fixed  $\alpha > 1$ . For the full optimization, we attach temporal upper indices, i.e., at time step  $t$ , we have the current approximate solution  $\eta^{(t)}$ , the Lagrange multipliers  $\lambda_l^{(t)}$  and the temperature parameter  $\tau^{(t)}$ . See Algorithm 1 for a description of the optimization scheme. While the number of optimization parameters grows quickly with the dimensionality of  $\theta$ , which may render the optimization challenging, in our experiments we did not encounter any issues with up to 54 optimization parameters and 40 constraints.

### D.3 Sampling from the Copula

A crucial step for our algorithms was baking the assumptions about  $p(X | Z)$  as well as  $Z \perp\!\!\!\perp U$  directly into our model from which we sample for Monte Carlo estimates. Algorithm 2 describes in detail how we can obtain these samples from the copula defined in eq. (4) in a differentiable fashion with respect to  $\eta$ .

### D.4 Parameter Initialization

We initialize the optimization parameters  $L$  with ones on the diagonal, zeros in the upper triangle, and sample the lower triangle from  $\mathcal{N}(0, 0.05)$ . The initialization for  $\mu_k$  and  $\ln(\sigma_k^2)$  depends on the chosen response function family. Our guiding principle is to ensure that the initial distribution covers a large set of possible response functions, tending towards larger  $\sigma_k$ .

## E Response Functions

One key advantage of our approach is that it allows us to flexibly trade off assumptions on the response function family with more informative bounds. Due to our simple, yet expressive choice of linear combinations of a set of basis functions, there are many natural and easy to implement options for the response functions. In particular, we consider the following options:

1. *Polynomials*:  $\psi_k(x) = x^{k-1}$  for  $k \in [K]$ . In this work, we specifically focus on linear ( $K = 2$ ), quadratic ( $K = 3$ ), and cubic ( $K = 4$ ) polynomial functions.
2. *Neural basis functions (MLP)*: We fit a multi-layer perceptron with  $K$  neurons in the last hidden layer to the observed data  $\{(x_i, y_i)\}_{i \in N}$  and take  $\psi_k(x)$  to be the activation of the  $k$ -th neuron in the last hidden layer. Note that the network output itself is a linear combination of these last hidden layer activations. Hence, the underlying assumption for this approach to work well is that the true causal effects can also be approximated well by a linear combination of the learned last hidden layer activations, i.e., the true effect is in this sense “similar” to the estimated observed conditional  $\hat{p}(y | x)$ . In practice, we train a 2-hidden layer MLP with 64 neurons in each layer,

rectified linear units as activation functions and an mean-squared-error loss for 100 epochs and a batchsize of 256 using Adam with a learning rate of 0.001.

3. *Gaussian process basis functions (GP)*: We fit a Gaussian process with a sum-kernel of a polynomial kernel of degree 3, an RBF kernel, and a white noise kernel to  $K$  different sub-samples  $\{(x_i, y_i)\}_{i \in N'}$  with  $N' \leq N$ . We then sample a single function from each Gaussian process as the basis functions  $\psi_k$  for  $k \in [K]$ . We train multiple Gaussian processes on smaller subsets of the data to ensure sufficient variance in the learned functional relation. Similarly to the neural net basis functions, the assumption is that the causal effect can be approximated by a linear combination of these varying samples. In our experiments, we fit the Gaussian processes with scikit-learn’s `GaussianProcessRegressor` (Pedregosa et al., 2011) using  $N' = 200$  and a white kernel variance of 0.4.

## F Why Discretization is not a Good Idea

The framework of Balke & Pearl (1994) is powerful and simple, and hence it raises the prospect that discretizing treatment  $X$  can provide a good approximation to the original problem where  $X$  is continuous. However, there are several reasons why this is not a good idea:

- *It destroys the key assumption of instrumental variable modeling.* Besides the lack of confounding between instrument  $Z$  and outcome  $Y$ , the key assumption in an IV model is the conditional independence  $Y \perp\!\!\!\perp Z \mid \{X, U\}$  (“exclusion restriction”). This assumption will in general fail to hold if we destroy information, i.e., if we condition on  $X \in \mathcal{A}$ , for some set  $\mathcal{A}$ , instead of the realization of  $X$ ;
- *It makes causal estimands ill-defined.* There are several ways in which an intervention can be ambiguous. This happens when defining the manipulation of a construct (“race”) or of summary measurements in general (“obesity”). One particular instance of the latter is when we speak of  $do(x^*)$ , meaning the setting of a discretization  $X^*$  of  $X$  to a particular level  $x^*$  (VanderWeele & Hernán, 2013). If  $X^* = x^*$  corresponds to the event  $X \in [a, b]$ , then this at least needs the assumption that  $\mathbb{E}[Y \mid do(x)]$  is approximately constant for  $x \in [a, b]$  for the intervention to be meaningful. This is pointless if the goal is to avoid making assumptions about the shape of the response function;
- *Its cost is super-exponential.* Suppose we still want to proceed with the idea of discretization, in the sense that we are willing to assume that we are using a fine enough grid of intervals for the treatment so that the previous two points are not particularly prominent. It may be argued that using Balke & Pearl (1994) with this approximation is attractive on the grounds it is a convex, deterministic approach and hence a more computationally attractive alternative to tackling the continuous problem. In fact, the opposite may hold. Assume we discretize  $X$  and  $Y$  to  $|\mathcal{X}|$  and  $|\mathcal{Y}|$  levels respectively, and  $Z$  assumes  $|\mathcal{Z}|$  levels (perhaps also by discretization). Then the cost of using the full information of the distribution is approximately  $\mathcal{O}(|\mathcal{X}|^{|\mathcal{Z}|}|\mathcal{Y}|^{|\mathcal{X}|})$ . It is true that, just like in our approach, this can be much simplified if we rely only on a subset of constraints. In particular, if we use only the first moments in the constraints and the expected outcome is the objective function, we can simplify the discrete formulation by targeting our parameterization to depend only on the expected outcomes directly. This makes the problem exponential only on  $|\mathcal{Z}|$ , see for instance the parameterization of Zhang & Bareinboim (2020). Being “only” exponential may still require Monte Carlo approximations in general. But this can still be super-exponential if  $|\mathcal{Z}|$  grows with  $|\mathcal{X}|$ , which will be necessary if the instrument is strong: for an extreme example, if  $Z$  and  $X$  lie close to a line with high probability and we choose only two levels of  $Z$  against many levels of  $X$ , then most combinations of pre-determined  $(z, x)$  pairs will lie on regions of essentially zero density in the  $p(x, z)$  distribution;
- *It is vacuous in the limit.* Even if we can use an arbitrarily fine discretization and assume that the piecewise nature of the approximation is close enough to the true response functions of  $Y$ , we know that as  $|\mathcal{X}| \rightarrow \infty$  the number of discontinuities in the response function also goes to infinity. As described by Günsilius (2018), we will not learn anything non-trivial about the causal estimand of interest.

We reiterate the points above in more direct way: *being unable to express constraints on the response function is not an asset, it’s a liability.* Discretization allows us to easily use a single family of functional constraints: piecewise constant functions. In this framework, it is cumbersome to represent other constraints such as smoothness constraints, and *the degree of violation of the exclusion*

*restriction assumption remains unknown.* There is no reason to believe this discrete representation is a good family in any computationally bounded sense, as an efficient choice of discretization points can only be made if we know something about the function. And if we do, then it makes far more sense to use more representationally efficient ways of partitioning the space of  $X$ , such as regression splines with a fixed number of knots. This involves no discretization of treatment, while avoiding the issues of violation of the exclusion restriction assumption and ambiguity of intervention.

## G Modeling $p(y | x, z)$

Alternatively to the setup described in the main text, we can match not only the marginal  $p(y | z)$ , but the theoretically more informative  $p(y | x, z)$ . This problem is actually conceptually simpler, although it will require joint measurements over the three types of variables.

The main modification is as follows. Instead of

$$\mathbb{E}[\phi_l(Y) | Z = z^{(m)}] = \int \phi_l(y_\theta(x)) p_\eta(x, \theta | Z = z^{(m)}) dx d\theta,$$

we build constraints based on

$$\mathbb{E}[\phi_l(Y) | X = x^{(m)}, Z = z^{(m)}] = \int \phi_l(y_\theta(x^{(m)})) p_\eta(\theta | X = x^{(m)}, Z = z^{(m)}) d\theta,$$

where now we need to define a grid over the joint space of  $X$  and  $Z$ . This can be done in several ways, including the joint product of equally-spaced quantiles of the respective marginal distributions, perhaps discarding combinations for which  $p(x^{(m)}, z^{(m)})$  are below some threshold. Moreover, the factor  $p_\eta(\theta | X = x^{(m)}, Z = z^{(m)})$  was explicitly parameterized in our original setup, and can be used as is.

Notice the advantages and disadvantages of the two approaches. Modeling the full conditional  $p(y | x, z)$  uses the full information of the problem (as it is equivalent to  $p(x, y | z)$ , where  $p(x | z)$  is tackled directly), which in principle is more informative but requires functionals of the joint  $p(x, y | z)$  instead of the marginals  $p(x | z)$  and  $p(y | z)$ . We can also see that we are trading-off adding more constraints but removing the need to integrate  $X$  in each constraint. More interestingly, this full conditional approach does not require any kind of density estimation: the need for  $p(x | z)$  disappears, and all we need on the left-hand sides are estimates of expectations.

## H Fitting Latent Variable Models

When fitting the latent variable models, we use multi-layer perceptrons with inputs  $z, x, y$  for the means and variances of the latent dimensions  $U$ , where we use lower indices  $U_i$  for the different components. For this encoder, we use 32 neurons in the hidden layer and rectified linear units as the activation function. There are two decoders. The first one is trained to reconstruct  $\mathbb{E}[X | X, U]$ , i.e., receives the original  $Z$  in addition to the latent vector  $U$  as input. It is also parameterized by an MLP with 32 neurons in the hidden layer and ReLU activations. The second decoder reconstructs  $\mathbb{E}[Y | X, U]$  and is either an MLP of the same architecture (when comparing to MLP response functions), linear in  $X$ , i.e.,  $\alpha X + \beta + \sum_{i=1}^{n_{\text{latent}}} (\gamma_i X U_i + \delta_i U_i)$  (when comparing to linear response functions), or quadratic in  $X$ , i.e.,  $\alpha X^2 + \beta X + \gamma + \sum_{i=1}^{n_{\text{latent}}} (\delta_i X^2 U_i + \epsilon_i X U_i + \zeta_i U_i)$  (when comparing to quadratic response functions). Thereby, we ensure that the form of matches our assumptions on the function form of the response family. We then optimize the evidence lower bound following standard techniques of variational autoencoders (Kingma & Welling, 2014) with  $L_2$  reconstruction loss for  $X$  and  $Y$ . We fit multiple models with different random initializations and compute the implied causal effect of  $X$  on  $Y$  for each one, which is obtained from the decoder  $\mathbb{E}[Y | u, x]$  by averaging over 1000 samples of the latent variable  $U$  for a fixed grid of  $x$ -values.

## I Additional Experimental Results

### I.1 Hyperparameter Settings

In all experiments, we fix hyperparameters  $M = 20$ ,  $L = 2$ ,  $B = 1024$  and run SGD with momentum 0.9 and learning rate 0.001 for 150 rounds of the augmented Lagrangian with 30 gradient updates

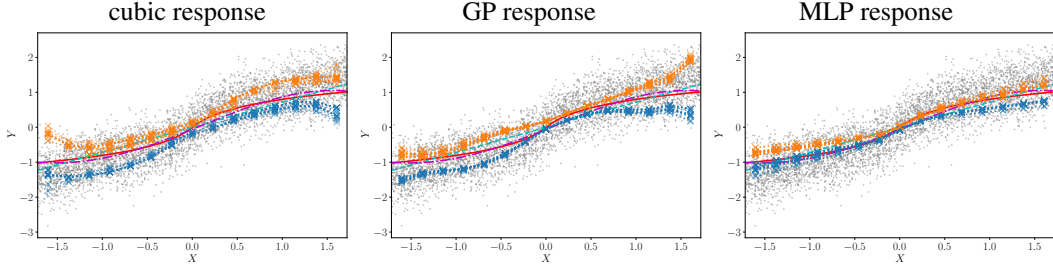


Figure 5: Bounds for the simulated sigmoidal design. The true causal effect is given by a logistic function, which is well recovered by our method for different response function families (cubic polynomials, GP basis functions, and MLP basis functions).

for each subproblem optimization. We start with a temperature parameter  $\tau = 0.1$  and multiply it by  $\alpha = 1.08$  in each round, capped at  $\tau_{\max} = 10$ . We use 7 neurons in the last hidden layer of the feed-forward neural net for MLP response functions in our synthetic setting and 9 for the expenditure data. For GP basis functions (see Appendix E), we sample 7 basis functions for the sigmoidal design dataset (see Appendix I.2). This set of hyperparameters did not require much manual tuning and worked for all datasets and response function families, i.e., also different dimensionality of  $\theta$ . For the synthetic settings, we sample 5000 observations each. We use 3 as the latent dimension when fitting our latent variable models. For the tolerances, we use  $\epsilon_{\text{abs}} = 0.2$  for the synthetic settings,  $\epsilon_{\text{abs}} = 0.1$  for the sigmoidal design (see Section I.2),  $\epsilon_{\text{abs}} = 0.3$  for the expenditure dataset and gradually tighten  $\epsilon_{\text{rel}}$  from 0.3 to 0.05 in all settings (which corresponds to the increasingly opaque lines).

## I.2 Sigmoidal Design

We also evaluate our method on simulated data from a sigmoidal design introduced by Chen & Christensen (2018), adopted by Newey & Powell (2003) and used in previous work on continuous instrumental variable approaches under the additive assumption as a common test case (Hartford et al., 2017; Singh et al., 2019; Muandet et al., 2020). We show the results from KIV and our bounds for response function families consisting of cubic polynomials and neural net basis functions in Figure 5. The observed data distribution  $\hat{p}(y|x)$  follows the true causal effect rather closely and the instrument is relatively strong in this setting, see Singh et al. (2019) for details. Therefore, the gap between our bounds is relatively narrow for a broad set of different basis functions as long as they are flexible enough to capture a sigmoidal shape.

## I.3 Expenditure Data

We prepare the data from Office for National Statistics (2000) using the same steps as Gunsilius (2020) closely following Newey & Powell (2003); Blundell et al. (2007). This is, we restrict the sample to households with married couples who live together and in which the head of the household is between 20 and 55 years old. We further exclude couples with more than 2 children. Finally, we also require the head of the household not to be unemployed. Otherwise, the instrument, gross earnings, would not be available. After these restrictions, we end up with 1650 observations in our dataset. The dataset can be downloaded for free for academic purposes after creating an account.

## I.4 Small Data Regime

Having tested our method on datasets of size 5000 (synthetic) and 1650 (expenditure data, see Appendix I.3), we now evaluate how our method performs on even smaller datasets. To this end, we first look at our synthetic settings using only 500 datapoints and correspondingly reducing the number of  $z$ -bins to  $M = 6$  in Figure 6. While the bounds are looser, our method can still provide useful information with relatively little data.

In addition, we ran our methods on a classic instrumental variable setting from economics, namely the dataset used by Acemoglu et al. (2001) on using settler mortality as an instrument to estimate

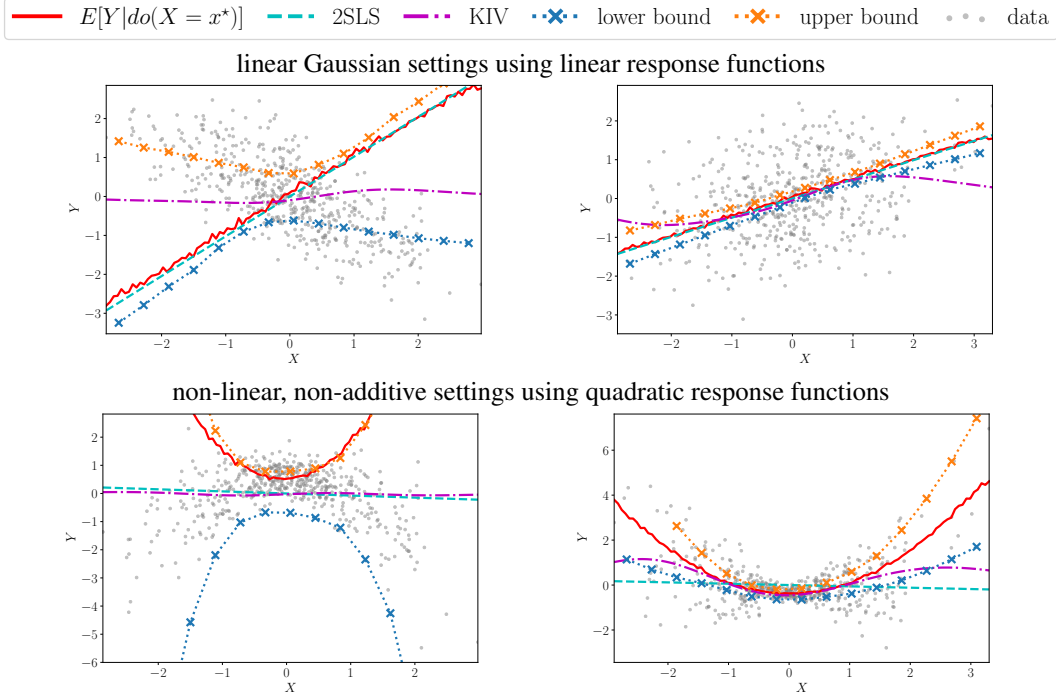


Figure 6: Performance of our method on smaller datasets with only 500 observations. The left column is the strong confounding weak instrument case ( $\alpha = 0.5, \beta = 3$ ) and the right column is the weak confounding strong instrument case ( $\alpha = 3, \beta = 0.5$ ).

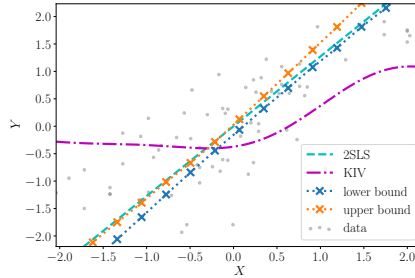


Figure 7: Results for the small dataset from Acemoglu et al. (2001) with linear response functions and  $M = 5$   $z$ -bins.

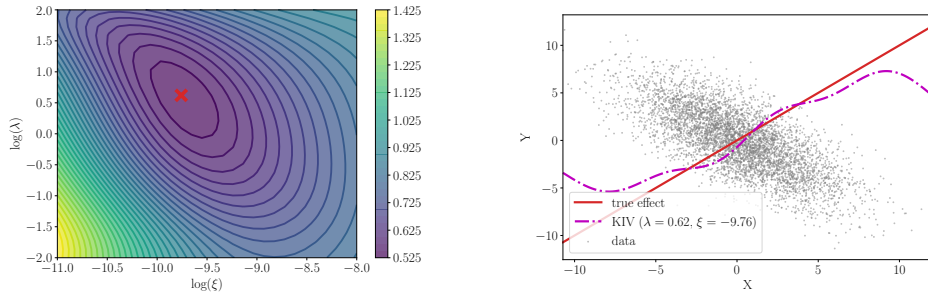
the causal effect of the health of institutions on economic performance.<sup>9</sup> This dataset consists of only 70 datapoints. Therefore, we set the number of  $z$ -bins to  $M = 5$  for this dataset. Restricting ourselves to linear response functions, our method still gives informative bounds, which include the effect estimated by 2SLS, but does not fully include the KIV results, see Figure 7.

## J KIV Heuristic for Tuning Hyperparameters

We have found KIV to fail in the strongly confounded linear Gaussian setting, even though all the assumptions are satisfied, see Figure 2 (row 1). Closer analysis of these cases showed that the heuristic that determines the hyperparameters does not return useful values in this setting. Instead, we performed a grid search over the main hyperparameters  $\lambda$  and  $\xi$  (see Singh et al., 2019, for details) and scored them by the out-of-sample mean-squared-error for the true causal effect (which is known in our synthetic setting). After manual exploration of the parameter space, we found a good setting

<sup>9</sup>The dataset is freely available at <https://economics.mit.edu/faculty/acemoglu/data/ajr2001>.

linear Gaussian setting with strong confounding and weak instrument ( $\alpha = 0.5, \beta = 3$ )



non-linear, non-additive setting with strong confounding and weak instrument ( $\alpha = 0.5, \beta = 3$ )

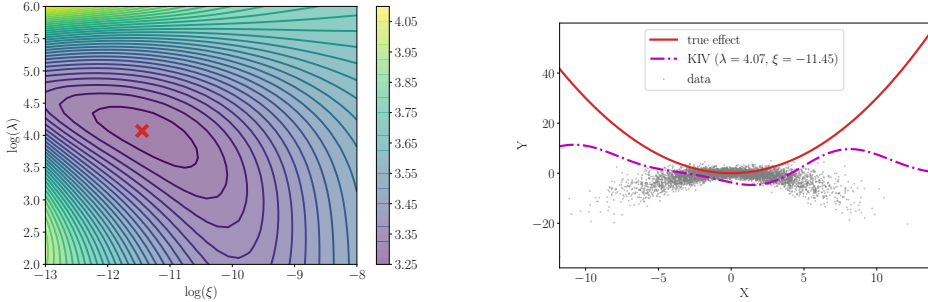


Figure 8: We show the results of a manual hyperparameter search for KIV in the left column, where we score different settings in the two-dimensional hyperparameter space by the log of the out-of-sample mean squared error, which requires knowledge of the true causal effect. The red cross denotes the setting with the smallest out-of-sample mean squared error. In the right column, we show the KIV regression lines using the hyperparameters found in the manual search. The first row corresponds to the linear Gaussian setting and the second row to the non-linear, non-additive synthetic setting.

marked by the red cross in the first row on the left of Figure 8. Using these fixed hyperparameters for KIV instead of the internal tuning stage, we get a much better approximation of the true causal effect shown in the first row on the right of Figure 8. Towards the data starved regions at large and small  $x$ -values, KIV again reverts back towards the prior mean of zero as expected. It is unclear at the moment, however, how to set such hyperparameter values without access to the true causal effect. Our point here is that in principle there is a setting with acceptable results, although even then it is not clear how much of it is a coincidence based on looking at many possible configurations.

We performed a similar manual analysis for the non-linear, non-additive synthetic setting with strong confounding, in which off-the-shelf KIV fails as well, see Figure 2 (row 3). Note that this setting does not satisfy the assumptions of KIV, because of the non-additive confounding. Again, we do manage to find hyperparameters that locally minimize the out-of-sample mean-squared-error shown in the second row on the left of Figure 8. However, the resulting regression of the causal effect does not properly capture the true effect as shown in the second row on the right of Figure 8.