BudgetIV: Optimal Partial Identification of Causal Effects with Mostly Invalid Instruments

Jordan Penn[1]   **Gecia Bravo-Hermsdorff**[3]   **Lee M. Gunderson**[2]
Ricardo Silva[2]   **David S. Watson**[1]
[1]King's College London   [2]University College London   [3]University of Edinburgh

Abstract

Instrumental variables (IVs) are widely used to estimate causal effects in the presence of unobserved confounding between an exposure X and outcome Y. An IV must affect Y *exclusively* through X and be *unconfounded* with Y. We present a framework for relaxing these assumptions with tuneable and interpretable "budget constraints". Our algorithm returns a feasible set of causal effects that can be identified exactly given perfect knowledge of observable covariance statistics. This feasible set might contain disconnected sets of possible solutions for the causal effect. We discuss conditions under which this set is *sharp*, i.e., contains all and only effects consistent with the background assumptions and the joint distribution of observable variables. Our method applies to a wide class of semiparametric models, and we demonstrate how its ability to select specific subsets of instruments confers an advantage over convex relaxations in both linear and nonlinear settings. We adapt our algorithm to form confidence sets that are asymptotically valid under a common statistical assumption from the Mendelian randomization literature. An accompanying R package, `budgetIVr`, is available from CRAN.

1   INTRODUCTION

The causal effect of a "treatment" (or "exposure") X on an "outcome" Y captures the change in the distribution of Y when we intervene on X (Pearl, 2009). In a randomized control trial, X is controlled explicitly so

that the causal effect can be identified (Neyman, 1923; Fisher, 1935). However, due to financial, ethical or physical constraints, confounding factors affecting X and Y often cannot be held fixed or observed and adjusted for (Rubin, 1974; Holland, 1986). In these situations, instrumental variables (IVs) are commonly used to infer causal effects.

IVs are a set of pre-treatment covariates $\boldsymbol{Z}$ that are (A1) associated with X; (A2) unconfounded with Y; and (A3) whose causal effect on Y is exclusively mediated through X (Angrist et al., 1996; Heckman and Vytlacil, 1999) (see Sect. 2.1 for formal definition). Due to the presence of latent confounders, one cannot test whether (A2) and (A3) hold for any particular pre-treatment covariate. However, with multiple candidate instruments, a simple fact may be exploited: if different sets of covariates predict different values of the causal parameter θ when treated as IVs, then at most one of the sets contains valid IVs.

Kang et al. (2016) exploit this fact by showing the following in linear models with a scalar treatment: if more than 50% of the pre-treatment covariates are correctly assumed to be valid IVs, then θ is point-identifiable (i.e., given an infinite number of observations from the model, the value of θ can be uniquely determined). A number of majority rule based approaches to inference with invalid IVs extend this work (Bowden et al., 2016a; Hartwig et al., 2017; Bucur et al., 2020; Hartford et al., 2021).

However, in settings where (a) 50% or fewer candidate IVs are valid, or (b) more than one causal parameter is needed (e.g., because the treatment is multidimensional), we show that θ can be at best *partially* identified. For example, even with oracle access to the joint distribution $P(X, Y, Z_1, Z_2)$, it may be undecidable whether Z_1 is a valid IV and Z_2 is not or vice versa. Though some values of θ may be excluded, the parameter will not converge to a single value even in the limit of infinite data.

Rather than assuming that all or a majority of $\boldsymbol{Z}$ are

valid IVs, our approach is to set a minimum proportion which are. In fact, we consider a more general setting (c), in which "degrees" of IV validity are set by user-input thresholds. It is not assumed *a priori* which thresholds apply to which candidate IVs. For each threshold, however, a user-input "budget" constrains how many candidates are allowed to violate the IV assumptions up to that degree. Bucur et al. (2020) and Xue et al. (2023) propose point estimators for θ under budget-like constraints, restraining fewer than 50% of candidate IVs to be valid. As we will see in Sect. 2, however, θ is only partially identified in this setting. The allocation of budgets to $\boldsymbol{Z}$ are also partially identified in general, which can lead to disconnected sets of solutions for θ.

Our discussion applies when IVs are invalid due to violation of (A2), (A3) or both. We show that under the assumption of homogeneous causal effects (Holland, 1986), which implies an additive separation of the $\boldsymbol{X}$- and $\boldsymbol{Z}$-signals into Y (Newey and Powell, 2003), a latent vector statistic summarizes bias in the estimated $\boldsymbol{\theta}$ due to violation of both assumptions. This statistic has a simple interpretation: the residual covariance between pre-treatment covariates $\boldsymbol{Z}$ and the outcome Y not explained by the causal effect of the treatment X. Vancak and Sjölander (2023) make a similar proposal to quantify the violation of IV assumptions in a single sensitivity parameter. However, their approach is limited to a single scalar Z, so does not utilize the notion of IV candidates and whether they give consistent estimates of θ.

We provide an algorithm, `budgetIV`, that provably finds *sharp* partial identification sets for causal effect parameters under fixed budget constraints. We show that the set can be inferred from relevant summary statistics, which reduce to the familiar covariance parameters $\text{Cov}(X, Z)$ and $\text{Cov}(Y, Z)$ for linear models. In the special case of scalar treatments, we provide a polytime procedure for computing sharp solution sets. However, under setting (b), we prove that partial identification is NP-hard in the number of instruments, budget thresholds and causal effect parameters. Finally, we provide a method for handling finite-sample uncertainty. In Sect. 4, we show how `budgetIV` can be adapted to form (asymptotically) valid confidence sets under the so-called "no measurement error" assumption from the Mendelian randomization literature (Bowden et al., 2016b).

The remainder of this paper is structured as follows. We formalize our problem in Sect. 2. We introduce the `budgetIV` algorithm in Sect. 3 and propose related inference procedures in Sect. 4. Experimental results are presented in Sect. 5. Following a literature review in Sect. 6, we conclude with a brief discussion in Sect. 7.
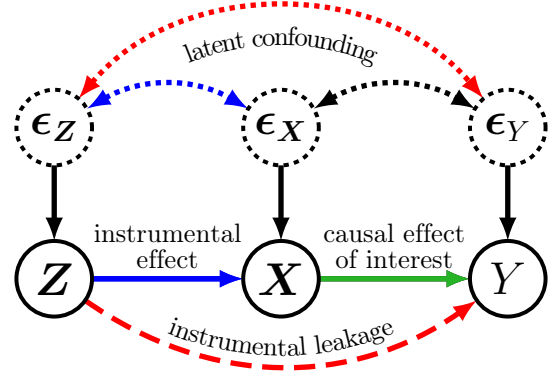

Figure 1: **Acyclic directed mixed graph for our problem setup.** Solid circles represent observable variables and dashed circles latent variables. Bidirected arrows are interpreted as any mutual dependence between noise residuals. The dotted black arrow indicates the unobserved confounding between $\boldsymbol{X}$ and Y. The relevance assumption (A1) requires at least one of the blue arrows. The red arrows contribute to violations of the exogeneity conditions (A2) and (A3). The green arrow indicates the causal effect of interest.

Proofs, pseudocode, and experimental details are given in the appendix.

2   PROBLEM SETUP

Our observable variables include a set of candidate instruments $\boldsymbol{Z} \in \Omega_{\boldsymbol{Z}} \subseteq \mathbb{R}^{d_{\boldsymbol{Z}}}$; treatments $\boldsymbol{X} \in \Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^{d_{\boldsymbol{X}}}$; and a univariate outcome $Y \in \Omega_Y \subseteq \mathbb{R}$. We assume that the ground truth structural equation model (SEM) between these variables takes the following form:

$$\boldsymbol{Z} := f_{\boldsymbol{z}}\left(\boldsymbol{\epsilon}_{\boldsymbol{z}}\right), \tag{1}$$

$$\boldsymbol{X} := f_{\boldsymbol{x}}\left(\boldsymbol{Z}, \boldsymbol{\epsilon}_{\boldsymbol{x}}\right), \tag{2}$$

$$Y := \boldsymbol{\theta}^* \cdot \boldsymbol{\Phi}(\boldsymbol{X}) + g_y(\boldsymbol{Z}, \boldsymbol{\epsilon}_{\boldsymbol{y}}), \tag{3}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{z}}$, $\boldsymbol{\epsilon}_{\boldsymbol{x}}$, and $\boldsymbol{\epsilon}_{\boldsymbol{y}}$ are noise residuals that capture any and all effects from latent variables, including unobserved confounding. The additive separability of $\boldsymbol{X}$ and $\boldsymbol{Z}$ in Y is guaranteed by assuming that the treatment effect is homogeneous, i.e., $P\big(y \mid \text{do}(\boldsymbol{x})\big) - P\big(y \mid \text{do}(\boldsymbol{x}_0)\big)$ is independent of $\boldsymbol{Z}$ and unobserved confounding between $\boldsymbol{X}$ and Y.

The function $\boldsymbol{\Phi} : \Omega_{\boldsymbol{X}} \mapsto \Omega_{\boldsymbol{\Phi}} \subseteq \mathbb{R}^{d_{\boldsymbol{\Phi}}}$ may provide a basis-expansion of a nonlinear treatment effect or a representation of a high-dimensional treatment $\boldsymbol{X}$. Though some authors have studied the underspecified regime in which $d_{\boldsymbol{\Phi}} > d_{\boldsymbol{Z}}$ (Pfister and Peters, 2022; Ailer et al., 2023), we restrict attention to the more common case where $d_{\boldsymbol{\Phi}} \leq d_{\boldsymbol{Z}}$. This guarantees the identification result of Thm. 1, first shown in linear

models by Koopmans (1949).

The vector $\boldsymbol{\theta^*} \in \mathbb{R}^{d_\Phi}$ is the causal parameter of interest. Together with $\boldsymbol{\Phi}$, it defines the average treatment effect:

$$\text{ATE}(\boldsymbol{x}; \boldsymbol{x}_0) = \boldsymbol{\theta^*} \cdot \big(\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x}_0)\big),$$

which represents the expected change in Y if we were to replace the intervention $do(\boldsymbol{X} = \boldsymbol{x}_0)$ with $do(\boldsymbol{X} = \boldsymbol{x})$.

2.1   A Sensitivity Parameter for IV Violation

We call $\boldsymbol{Z}$ a set of valid IVs if they satisfy the following properties:

(A1) *Association*: $\boldsymbol{Z} \not\!\perp\!\!\!\perp \boldsymbol{X}$,
(A2) *Unconfoundedness*: $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{\epsilon_y}$,
(A3) *Exclusion*: $\boldsymbol{Z} \perp\!\!\!\perp Y \mid \{\boldsymbol{X}, \boldsymbol{\epsilon_y}\}$.

In several well-studied regimes (Koopmans, 1949; Imbens and Angrist, 1994; Newey and Powell, 2003; Heckman and Vytlacil, 2005; Saengkyongam et al., 2022), the assumptions above allow for point identification of the target parameter $\boldsymbol{\theta^*}$. For our SEM (Eqs. (1) to (3)), we prove point identification under a stronger version of (A1) and an assumption that combines aspects of (A2) and (A3), despite being strictly weaker than either. To show this, we introduce the following parameters:

$$\begin{aligned}
\boldsymbol{\gamma_g} &:= \text{Cov}\big(g_y(\boldsymbol{Z}, \boldsymbol{\epsilon_y}), \boldsymbol{Z}\big), \\
\boldsymbol{\beta_\Phi} &:= \text{Cov}\big(\boldsymbol{\Phi}(\boldsymbol{X}), \boldsymbol{Z}\big), \\
\boldsymbol{\beta_y} &:= \text{Cov}\big(Y, \boldsymbol{Z}\big).
\end{aligned}$$

By taking the covariance between each term in Eq. (3) with $\boldsymbol{Z}$, we see that target parameter $\boldsymbol{\theta^*}$ is related to $\boldsymbol{\gamma_g}$ by $\boldsymbol{\gamma_g} = \boldsymbol{\beta_y} - \boldsymbol{\theta^*} \cdot \boldsymbol{\beta_\Phi}$. Thus, $\boldsymbol{\gamma_g}$ is the residual covariance between Y and $\boldsymbol{Z}$ not explained by the ground truth causal effect of $\boldsymbol{X}$ on Y. Notice that if both (A2) and (A3) are satisfied, then $\boldsymbol{\gamma_g} = \boldsymbol{0}$.

Given that $d_\Phi \leq d_{\boldsymbol{Z}}$, the following constraints are sufficient for identification:

(B1*) *Association (strong)*: $\text{rank}(\boldsymbol{\beta_\Phi}) = d_\Phi$
and $(\boldsymbol{\beta_\Phi})_i \neq 0$ ($\forall Z_i \in \boldsymbol{Z}$);
(B2*) *Exogeneity (strong)*: $\boldsymbol{\gamma_g} = \boldsymbol{0}$,

where $(\boldsymbol{\beta_\Phi})_i := \text{Cov}\left(\boldsymbol{\Phi}(\boldsymbol{X}), Z_i\right)$ and "exogeneity" refers to the conjunction of (A2) and (A3).

Theorem 1 (Identifiability). Assume Eqs. (1) to (3) and claims (B1*), (B2*) hold for some $d_\Phi \leq d_{\boldsymbol{Z}}$. Assume the existence of a ground truth joint distribution $P(\boldsymbol{X}, Y, \boldsymbol{Z})$ with finite covariance parameters $\boldsymbol{\beta_\Phi^*}, \boldsymbol{\beta_y^*}$. Then the causal parameter $\boldsymbol{\theta^*}$ can be identified exactly as the unique solution to $\boldsymbol{\beta_y^*} - \boldsymbol{\theta^*} \cdot \boldsymbol{\beta_\Phi^*} = \boldsymbol{0}$.

We can relax (B1*) and (B2*) further, in particular using $\boldsymbol{\gamma_g}$ to model violations of exogeneity:

(B1) *Association (relaxed)*: $\boldsymbol{\beta_\Phi} \neq \boldsymbol{0}$;
(B2) *Exogeneity (relaxed)*: $\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}$.

Such a relaxation may still allow for partial identification, which we will see in the following section. In Sect. 2.3 we introduce budget constraints as tuneable and interpretable choices for $\boldsymbol{\Gamma}$.

2.2   Formalizing Optimal Partial Identification

We say that the causal parameter $\boldsymbol{\theta}^*$ is *partially identified* when more than one (but not all) of its possible values are consistent with the data and our structural assumptions. In this section, we define a certain notion of *optimality* for general partial identification problems, and establish the ingredients for an optimal solution in our setting.

The following definitions allow us to state and prove these results. Let $\mathcal{M}$ be a class of SEMs and $m^* \in \mathcal{M}$ the ground truth model. Each $m \in \mathcal{M}$ implies a joint distribution $P_m(\boldsymbol{V})$ over the observables $\boldsymbol{V} \in \Omega_{\boldsymbol{V}}$. (In our setting $\boldsymbol{V} = \{\boldsymbol{X}, Y, \boldsymbol{Z}\}$ and $\Omega_{\boldsymbol{V}} = \Omega_{\boldsymbol{X}} \times \Omega_Y \times \Omega_{\boldsymbol{Z}}$.)

A constraint $c : \mathcal{M} \mapsto \{0, 1\}$ is a logical formula that either does or does not hold for any given model. For instance, c may bound the range of some parameter(s) in m or impose conditional independence on certain variables in $\boldsymbol{V}$. Let $\mathcal{C}$ be a set of such constraints, with $\mathcal{C}^* := \{c \in \mathcal{C} : c(m^*) = 1\}$ denoting ground truth. (In our setting, $\mathcal{C}$ includes the relaxed association and exogeneity assumptions (B1) and (B2).)

An *observable* statistic s is a functional of the joint distribution, $s : \{P_m(\boldsymbol{V}) : m \in \mathcal{M}\} \mapsto \Omega_s$, with ground truth value $s^* := s\left(P_{m^*}(\boldsymbol{V})\right)$. Examples include (conditional) moments or correlations between variables. (In our setting, s comprises the cross-covariance parameters $\boldsymbol{\beta_y}, \boldsymbol{\beta_\Phi}$.) The target parameter $q^* := q(m^*)$ is the ground truth for some *latent* statistic $q : \mathcal{M} \mapsto \Omega_q$, which cannot be determined by $P_m(\boldsymbol{V})$ alone. (In our setting, q^* is the causal parameter $\boldsymbol{\theta}^*$.)

Given a constraint $c \in \mathcal{C}$ and statistic s, the "plausible" values of q^* form a *solution set* $\mathcal{T}(c, s) \subseteq \Omega_q$. Such sets are the image of a *solution map* for q, $\mathcal{T} : \mathcal{C} \times \Omega_s \mapsto \mathcal{P}(\Omega_q)$, where $\mathcal{P}$ denotes the power set.

We define an *optimal* solution map in terms of soundness, completeness, and minimality criteria. We defer discussion of computational complexity and finite sample inference to Sect. 3 and 4, respectively.

Definition 1 (Soundness). A solution map $\mathcal{T}$ is *sound* if, for any ground truth model $m^* \in \mathcal{M}$, given

statistic $s^* := s(P_{m^*}(\boldsymbol{V}))$ and constraint $c^* \in \mathcal{C}^*$, we have $q^* \in \mathcal{T}(c^*, s^*)$.

This condition ensures that our solution map cannot exclude the target q^* when provided with ground truth inputs.

Definition 2 (Completeness). A solution map $\mathcal{T}$ is *complete* if, for any ground truth model $m^* \in \mathcal{M}$, given s^* and any $c^* \in \mathcal{C}^*$, the following holds. For all $q \in \mathcal{T}(c^*, s^*)$, there is at least one model $m_q \in \mathcal{M}$ for which $q = q(m_q)$ and $s(P_{m_q}(\boldsymbol{V})) = s^*$.

This condition ensures that no sound map can exclude more values of $q \in \Omega_q \backslash \{q^*\}$. Solution sets that are both sound and complete are said to be *sharp*, as originally defined by Manski (1990, 2003).

For the next definition, we introduce a partial order on observable statistics with respect to their information content. We say that s' is at least as informative as s if there exists a deterministic function $f : \Omega_{s'} \mapsto \Omega_s$ such that, for any $m \in \mathcal{M}$, $s(P_m(\boldsymbol{V})) = f(s'(P_m(\boldsymbol{V})))$. In this case, we write $s' \preceq s$.

Definition 3 (Minimality). A sharp solution map $\mathcal{T} : \mathcal{C} \times \Omega_s \mapsto \Omega_q$ is *minimal* if, for any other sharp solution map $\mathcal{T}' : \mathcal{C} \times \Omega_{s'} \mapsto \Omega_q$ that takes a different input statistic s', we have $s' \preceq s$.

Minimality ensures that no sharp map could be constructed using strictly less information. With Defs. 1 to 3 in place, we can show the existence of an optimal solution map for our problem.

Theorem 2 (Optimal solution map). Let the model class $\mathcal{M}$ consist of all SEMs consistent with Eqs. (1) to (3) for some given $\boldsymbol{\Phi}$ and $d_{\boldsymbol{\Phi}} \leq d_{\boldsymbol{Z}}$. Assume the existence of a ground truth joint distribution $P(\boldsymbol{X}, Y, \boldsymbol{Z})$ with finite covariance parameters $\boldsymbol{\beta}_{\boldsymbol{\Phi}}^*, \boldsymbol{\beta}_{\boldsymbol{y}}^*$. Define the affine function $h(\boldsymbol{\theta}) := \boldsymbol{\beta}_{\boldsymbol{y}} - \boldsymbol{\theta} \cdot \boldsymbol{\beta}_{\boldsymbol{\Phi}}$, for some $\boldsymbol{\beta}_{\boldsymbol{y}}, \boldsymbol{\beta}_{\boldsymbol{\Phi}}$. Then the solution map $\mathcal{T}$ defined by:

$$\mathcal{T}(c, s = (\boldsymbol{\beta}_{\boldsymbol{\Phi}}, \boldsymbol{\beta}_{\boldsymbol{y}})) = \{\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\Phi}}} : h(\boldsymbol{\theta}) \in \boldsymbol{\Gamma}\}, \quad (4)$$

is sound, complete, and minimal with respect to background constraint set $\mathcal{C} = \{\mathbb{I}[\boldsymbol{\gamma}_{\boldsymbol{g}} \in \boldsymbol{\Gamma}] : \boldsymbol{\Gamma} \subseteq \mathbb{R}^{d_{\boldsymbol{Z}}}\}$, where $\mathbb{I}[\cdot]$ denotes the indicator function.

In Appx. B.1 we examine how extra assumptions can result in smaller solution sets. In particular, we show that this is the case for models with categorical (Balke and Pearl, 1997) or bounded (Manski, 1990) outcomes.

2.3 Budget Background Constraints

We introduce sensible choices for the search space $\boldsymbol{\Gamma}$. For any $\boldsymbol{\gamma}_{\boldsymbol{g}} \in \mathbb{R}^{d_{\boldsymbol{Z}}}$ and any $\alpha \in [0, 1)$, if $\boldsymbol{\gamma}_{\boldsymbol{g}}' := \alpha \boldsymbol{\gamma}_{\boldsymbol{g}}$, the degree of IV violation attributed to each candidate and the combined violation is strictly weaker for $\boldsymbol{\gamma}_{\boldsymbol{g}}'$

than for $\boldsymbol{\gamma}_{\boldsymbol{g}}$. Below, we formalize the notion that if $\boldsymbol{\gamma}_{\boldsymbol{g}}$ is plausible, then so too is $\boldsymbol{\gamma}_{\boldsymbol{g}}'$.

Definition 4 (Star domain). A set of points $A \subseteq \mathbb{R}^d$ is a *star domain* if there exists a point $\boldsymbol{a} \in A$ such that the line between $\boldsymbol{a}$ and any other point $\boldsymbol{a}' \in A$ is contained within A. Equivalently, a star domain is a space $A \subseteq \mathbb{R}^d$ with a nonempty convex kernel:

$$\text{ck}(A) := \{\boldsymbol{a} \in A \mid \forall \boldsymbol{a}' \in A, \eta \in [0, 1] : \\ \eta \boldsymbol{a} + (1 - \eta) \boldsymbol{a}' \in A\}.$$

Principle 1 (Starfish principle). Relax a structural assumption by adding a latent sensitivity parameter, $\boldsymbol{\delta}$, whose direction is related to the mechanisms of violation and whose magnitude increases with the degree of violation by these mechanisms. Bound the support of this parameter within a star domain for the FeasIble SearcH space—i.e., a starFISH—whose convex kernel includes the point $\boldsymbol{\delta} = \boldsymbol{0}$.

This principle encapsulates current literature restricting the number[1] (Kang et al., 2016; Hartwig et al., 2017; Silva and Shimizu, 2017; Hartford et al., 2021) or total effect (Ramsahai, 2012; Conley et al., 2012; Silva and Evans, 2016; Watson et al., 2024; Jiang and Kocaoglu, 2024) of invalid instruments. Convex restrictions lead to convex identified sets when the sensitivity parameter is an affine function of the causal effect parameters. If this affine function intersects the convex kernel, the intersection is connected (by definition). In the general case, the identified set for the quantities of interest may be disconnected (see Fig. 2). However, each disjoint subset might also provide a distinct mechanistic interpretation for the plausible violations.

Here, we introduce budget constraints, which are based on two user-specified components: thresholds, $\boldsymbol{\tau} := (\tau_1, \tau_2, \ldots, \tau_K)$, which describe degrees of IV invalidity; and budgets, $\boldsymbol{b} := (b_1, b_2, \ldots, b_K)$, which determine the minimum number of IVs assumed to be at least as valid as the corresponding threshold. The thresholds must be nonnegative and increasing: $0 \leq \tau_1 < \tau_2 < \ldots < \tau_K < \infty$. Integer budgets are increasing and strictly positive: $0 < b_1 < b_2 < \cdots < b_K \leq d_{\boldsymbol{Z}}$. Thresholds and budgets with the same index $i \in [K]$ form pairs. The number of thresholds–budget pairs K is chosen by the user but we require $K \leq d_{\boldsymbol{Z}}$.

We define the $d_{\boldsymbol{Z}}$-dimensional latent statistic $\boldsymbol{U}(\boldsymbol{\gamma}_{\boldsymbol{g}})$ through the following relationship:

$$U_i = \ell \iff \tau_{\ell-1} \leq |\gamma_{g_i}| \leq \tau_\ell.$$

Let $U_{i1}, U_{i2}, \ldots, U_{iK}$ represent the encoding of U_i such that $U_{i\ell} = \mathbb{I}[U_i \leq \ell]$ for all $\ell \in [K]$, where $\mathbb{I}[\cdot]$ is the

---

[1]Notice that the L_0 norm constraint $\|h(\boldsymbol{\theta})\|_0 \leq b$ is a star domain with a convex kernel including the point $\boldsymbol{\theta} = \boldsymbol{0}$.
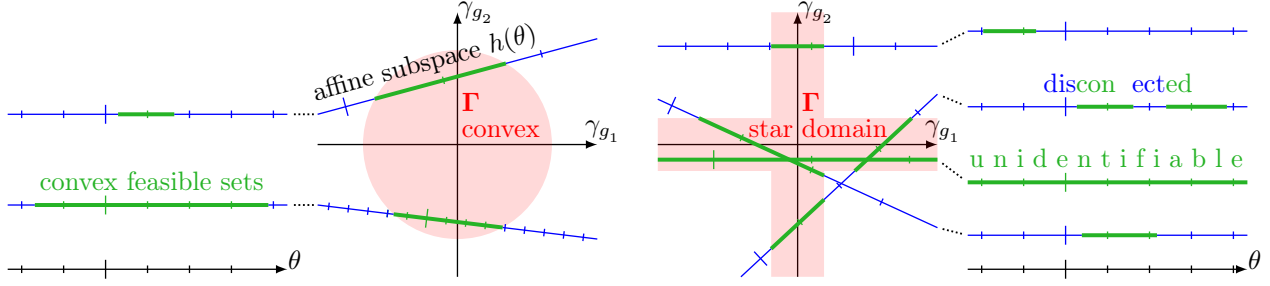
Figure 2: **The topology of a feasible set depends on the shape of the background constraints.**
Plots of constrained search spaces $\mathbf{\Gamma}$ (shaded) and lines $h(\theta)$ corresponding to $d_{\mathbf{Z}} = 2, d_{\Phi} = 1$. The intersection between a line and shaded region determines a feasible set of θ. The constraints are a subspace of $\mathbb{R}^{d_{\mathbf{Z}}}$ while $h(\theta)$ are d_{Φ}-dimensional affine subspaces. (Left) Convex $\mathbf{\Gamma}$ entails convex feasible sets. (Right) Budget constraints $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ form a star domain. They are non-convex in general and are unbounded for $b_K < d_{\mathbf{Z}}$ (i.e., some γ_{g_i} may be unconstrained). This can lead to disconnected or even unidentifiable causal effect. In Appx. B.2 we show that unidenfitiability occurs only under violation of (B1*) and can be tested in polytime.

indicator function. We define the budget constraint restriction $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ as the set of vectors $\boldsymbol{\gamma_g}$ satisfying, for all $\ell \in [K], \sum_{i=1}^{d_{\mathbf{Z}}} U_{i\ell} \geq b_\ell$ for some choice of budgets and thresholds. We denote the set of $\boldsymbol{U}$ with encodings satisfying the inequality above by $\boldsymbol{\Sigma_b}$. In particular, if equality holds, we say $\boldsymbol{U} \in \boldsymbol{\Sigma_b}^{(\max)}$ is a maximally relaxed assignment. These definitions will be useful in the next section.

Since $\boldsymbol{U}(\boldsymbol{\gamma_g})$ depends on $\boldsymbol{\theta}^*$ it is only partially identifiable in general. We define the set of plausible budget assignments:

$$\{\boldsymbol{U}(h(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \mathbb{R}^{d_{\mathbf{Z}}}, h(\boldsymbol{\theta}) \in \mathbf{\Gamma}(\boldsymbol{b}, \boldsymbol{\tau})\}.$$

For instance, if $K = 1$ and $\tau_1 = 0$, we assume at least b_1 candidate IVs are valid (having $\gamma_{g_i} = 0$), but allow $d_{\mathbf{Z}} - b_1$ to be unrestricted. This extends the L_0-norm constraint from Kang et al. (2016) to violation of (A2) and/or (A3).

3   THE ALGORITHM

The objective of budgetIV is twofold: (O1) discover plausible budget assignments; and (O2) partially identify the ATE. We simplify (O1), which for general $d_{\mathbf{Z}} > d_{\Phi} > 1$ allows the solution set for $\boldsymbol{\theta}$ to be calculated using subset selection over $\boldsymbol{\Sigma_b}^{(\max)}$. We conjecture that subset selection cannot be avoided, which means finding any $\boldsymbol{\theta}$ in the solution set is NP-hard (w.r.t. $d_{\mathbf{Z}}$ and d_{Φ}), while finding the entire solution set is #P-hard. As it stands, budgetIV should not be run for $d_{\mathbf{Z}}, d_{\Phi} > 10$. When $d_{\Phi} = 1$, however, we show the solution set can be found in polynomial time using our algorithm budgetIV_scalar. Pseudocode for both algorithms is given in Appx. C; executable code is available online.[2]

Simplifying (O1) **for** $d_{\Phi} > 1$. Since $\boldsymbol{U}(\boldsymbol{\gamma})$ is uniquely defined for each $\boldsymbol{\gamma} \in \mathbb{R}^{d_{\mathbf{Z}}}$, the feasible region $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ decomposes into subsets $\mathbf{\Gamma_U}$ for which $\boldsymbol{\gamma} \in \mathbf{\Gamma_U} \implies \boldsymbol{U}(\boldsymbol{\gamma}) = \boldsymbol{U}$. However, in Appx. E, we show that $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ can also be thought of as a union of overlapping cuboids. Indeed, in Fig. 2 we see a feasible region made up of two overlapping rectangles. Each cuboid $\tilde{\mathbf{\Gamma}}_{\tilde{\boldsymbol{U}}}$ is indexed by a maximal budget assignment $\tilde{\boldsymbol{U}} \in \boldsymbol{\Sigma_b}^{(\max)}$.

3.1   budgetIV with Oracle β Parameters

We summarize the budgetIV algorithm below.

(1) For each $\tilde{\boldsymbol{U}} \in \boldsymbol{\Sigma}^{(\max)}$, test for the intersection between $h(\boldsymbol{\theta}) = \boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi}$ and $\tilde{\mathbf{\Gamma}}_{\tilde{\boldsymbol{U}}}$.

(2) If there is an intersection, solve the following linear program to find the bounds on $\text{ATE}(\boldsymbol{x}; \boldsymbol{x}_0)$ for each $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ of interest and baseline treatment $\boldsymbol{x_0}$:

$$\min_{\boldsymbol{\theta}: \, h(\boldsymbol{\theta}) \in \tilde{\mathbf{\Gamma}}_{\tilde{\boldsymbol{U}}}} /\max \quad \boldsymbol{\theta} \cdot \big(\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x}_0) \big).$$

(3) Let $\boldsymbol{\theta}^{-/+}(\boldsymbol{x})$ denote the argmin/argmax solution (respectively) to the optimization problem in step (2) at point $\boldsymbol{x}$. The function $U_i(h(\boldsymbol{\theta}))$ returns the value of the latent variable U_i at the point $h(\boldsymbol{\theta})$. Compute $\check{\boldsymbol{U}}$, defined by its components:

$$\check{U}_i := \max_{\boldsymbol{x} \in \Omega_{\boldsymbol{X}}} \max_{\diamond \in \{-,+\}} U_i\big(\boldsymbol{\theta}^{\diamond}(\boldsymbol{x})\big).$$

(4) Return all unique $\check{\boldsymbol{U}}$ (O2) along with the corresponding ATE bounds (O1).

Polytime budgetIV_scalar. The algorithm above describes an ILP that requires a linear search over $|\Sigma_{\boldsymbol{b}}^{(\max)}|$, which itself is bounded above by $d_{\boldsymbol{Z}}^{d_{\boldsymbol{Z}}}$, corresponding to the case where $K = d_Z$. Though standard

solvers are highly optimized, this task can quickly become intractable with many candidate instruments. When $d_\Phi = 1$, however, exact partial identification of the ATE can be computed in $\tilde{\mathcal{O}}(d_Z K)$ time. This is done by noticing that the value of U associated with $h(\theta)$ can only change as a function of θ at the $2 d_Z K$ points:

$$\Theta_{ik}^{\pm} := \frac{\pm \tau_k - (\beta_y)_i}{(\beta_\Phi)_i},$$

where we ignore any i for which $(\beta_\Phi)_i = 0$. Similarly, the full set of plausible budget assignments can be calculated in polynomial time. See Appx. C for further details.

3.2 The L_0-norm Constraint

Kang et al. (2016) show that point identification is possible in the linear setting with univariate exposure X and some invalid instruments—provided at least half are valid. We show that this result cannot be generalized to multidimensional exposures. However, we place tight bounds on the cardinality of the feasible set with respect to the minimum number of valid IVs b. We use the following shorthand:

$$\mathcal{T}_{L_0}(b) := \mathcal{T}\left(c = \mathbb{I}[\gamma_g \in \Gamma(0, b)], s = \{\beta_\Phi, \beta_y\}\right).$$

Theorem 3 (t-point identification). Assume Eqs. 1, 2, 3 and claims (B1*), (B2) hold for some $\Gamma(0, b)$. Then, for all $b < d_Z - d_\Phi + 1$, the cardinality of the feasible set $t := |\mathcal{T}_{L_0}(b)| \in \mathbb{N}$ is bounded above by:

$$t \le \frac{d_Z!}{(d_\Phi - 1)!(d_Z - d_\Phi + 1)!} \left\lfloor \frac{d_Z - d_\Phi + 1}{d_Z - b - d_\Phi + 1} \right\rfloor.$$

This bound is tight in the sense that equality holds for some values of β_Φ, β_y. We extend these results to incorporate violation of (B1*) in the proof (Appx. A.3).

Corollary 3.1 (No point identification for $d_\Phi > 1$). There is no value of b for which $t \le 1$ is guaranteed for all $|\mathcal{T}_{L_0}|$ when $d_\Phi > 1$.

Corollary 3.2 (t-point identification for $d_\Phi = 1$). In the case of $d_\Phi = 1$, we have:

$$|\mathcal{T}_{L_0}(b)| =: t \le \left\lfloor \frac{d_Z}{d_Z - b} \right\rfloor \le d_Z.$$

This reduces to point identifiability when $b < d_Z / 2$.

4 INFERENCE

In this section we consider finite-sample uncertainty. We seek confidence sets with (asymptotically) valid coverage over the solution set $\{\theta \in \mathbb{R}^{d_Z} : h(\theta) \in \Gamma(\tau, b)\}$.

Since $\mathcal{T}$ is a deterministic map, it is possible to use a confidence set (B_Φ, B_y) over the covariance parameters β_Φ, β_y. The confidence set over the causal parameter is then all $\theta \in \mathbb{R}^{d_\Phi}$ satisfying:

$$\forall (\beta_\Phi, \beta_y) \in (B_\Phi, B_y) : (\beta_y - \theta \cdot \beta_\Phi) \in \Gamma(\tau, b).$$

Provided the estimators $\hat{\beta}_\Phi, \hat{\beta}_y$ have finite variance, an asymptotically valid confidence set can be constructed by modeling these estimators as multivariate normal (see Appx. A.4). However, we choose to make the following simplifications taken from the applied IV literature.

4.1 Coverage with Summary Statistics under the NOME Assumption

Candidate IVs are often selected from a pool of covariates, with inclusion based solely on marginal association with the exposure. For instance, Mendelian randomization (MR) is a popular approach in genetic epidemiology whereby genetic variants Z are used as IVs to determine the causal effect of phenotype(s) X on a health outcome Y. The $Z \to X$ link is usually established by a genome-wide association study (GWAS). Empirically, the chosen Z tend to be less strongly associated with Y than with X. This can occur for various reasons: strong latent confounding between X and Y; a weak causal effect of X on Y; low variation in Y (e.g., rare diseases); and/or smaller sample sizes for evaluating $P(Z, Y)$ than $P(Z, X)$ (Pierce and Burgess, 2013).

In such cases, finite-sample error in $\hat{\theta}$ is mostly explained by finite-sample error in $\hat{\beta}_y$. This has led to the introduction of a *no measurement error* (NOME) assumption in MR studies (Bowden et al., 2016b, 2018), under which one assumes finite sample error is only due to error in $\hat{\beta}_y$. For complex choices of Φ, it may also be the case that p-values for testing $(\hat{\beta}_\Phi)_{ij} = 0$ are lower than those for testing $(\hat{\beta}_y)_i = 0$, which would justify the NOME assumption. We have adapted budgetIV to construct confidence sets for θ with (asymptotically) valid probabilities to cover the entire solution set $\{\theta \in \mathbb{R}^{d_\Phi} : h(\theta) \in \Gamma(\tau, b)\}$ under the NOME assumption.

In particular, we use a Bonferroni adjustment to construct a box-shaped confidence set over β_y. For a target coverage $(1 - \alpha) \times 100\%$, we take a union over $(1 - \alpha/d_Z) \times 100\%$ confidence intervals corresponding to each $(\beta_y)_j$. Let $(\delta \beta_y)_i$ denote the half width of the confidence interval over $(\beta_y)_i$. While box-shaped confidence sets are conservative, they neatly superimpose
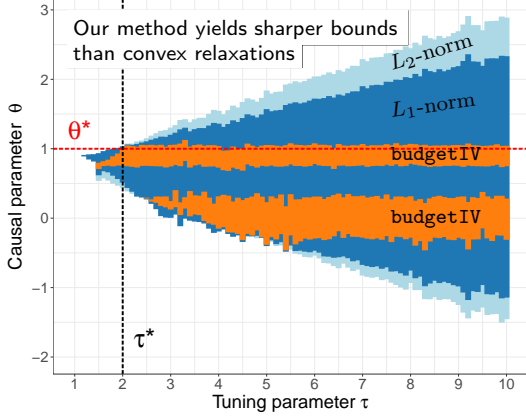
Figure 3: **Our method yields sharper bounds than convex relaxations in linear models.** Bounds on θ for a series of linear models with scalar exposure X and $\gamma_g^* = (-2, -0.4)$. Plug-in estimators are used throughout. Orange bounds come from `budgetIV` with $\tau = (0.6, \tau)$; dark blue from the L_1-norm constraint $\|\gamma_g\|_1 \leq \tau + 0.6$; and light blue from the L_2-norm constraint $\|\gamma_g\|_2 \leq \sqrt{\tau^2 + 0.6^2}$. We vary τ linearly from 0 to 10—each bound is an experiment.
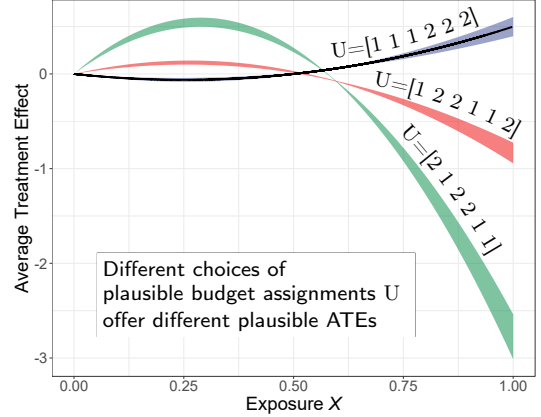
Figure 4: **Budget constraints provide information about the structure of the problem.** Feasible values of the ATE relative to a baseline of $x_0 = 0$ as exposure X varies in a nonlinear SEM with $d_{\mathbf{Z}} = 6$. The true ATE is given by the solid black curve. Each colored region corresponds to a unique intersection of γ_g and the star domain $\mathbf{\Gamma}$. The union of such intersections at each value of X produces a disconnected feasible set.

with the τ-thresholds:

$$\left| (\hat{\beta}_y)_i \pm (\delta\beta_y)_i - (\boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}}_{\mathbf{\Phi}})_i \right| \leq \tau$$
$$\iff \left| (\hat{\beta}_y)_i - (\boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}}_{\mathbf{\Phi}})_i \right| \leq \tau + (\delta\beta_y)_i.$$

We use this relationship to, we add slack $(\pm \delta\beta_y)_i$ to the corresponding face of each $\tilde{\mathbf{\Gamma}}_{\tilde{\mathbf{U}}}$ for $\tilde{\mathbf{U}} \in \mathbf{\Sigma}^{(\max)}$ to form $\hat{\mathbf{\Gamma}}_{\tilde{\mathbf{U}}}$, and optimize over these domains to construct corresponding confidence intervals for the $\text{ATE}(\boldsymbol{x}; \boldsymbol{x_0})$ (see Sect. 3.1). Our approach has the following guarantee.

Theorem 4 (Coverage). Fix the target level $\alpha \in (0, 1)$. Let $\hat{\mathbf{\Gamma}}_\alpha := \bigcup_{\tilde{\mathbf{U}} \in \mathbf{\Sigma}^{(\max)}} \hat{\mathbf{\Gamma}}_{\tilde{\mathbf{U}}}$ be formed from the $(1 - \alpha/d_{\mathbf{Z}}) \times 100\%$ confidence intervals for each component of $\hat{\boldsymbol{\beta}}_y$ as described above, where the intervals are estimated from a dataset of n samples drawn iid from $P(\boldsymbol{X}, Y, \boldsymbol{Z})$. Using the shorthand:

$$\hat{\mathcal{T}}_\alpha := \mathcal{T}\left(c = \mathbb{I}[\gamma_g \in \hat{\mathbf{\Gamma}}_\alpha], s = (\boldsymbol{\beta}_{\mathbf{\Phi}}^*, \hat{\boldsymbol{\beta}}_y) \right),$$

we have, as $n \to \infty$:

$$P\left(\boldsymbol{\theta}^* \in \hat{\mathcal{T}}_\alpha \right) \geq 1 - \alpha.$$

5 EXPERIMENTS

The full simulation studies are detailed in Appx. D, along with additional results.

Sharper than convex relaxations. In Fig. 3 we study the effect of varying the thresholds $\boldsymbol{\tau}$ on the feasible set. We consider a linear Gaussian model with $d_X = 1$ and $d_{\mathbf{Z}} = 2$ where (A2) is violated through correlation between $\boldsymbol{Z}$ and ϵ_y, while (A3) is satisfied. We fix the ground truth parameters $\theta^* = 1$, $\gamma_g^* = (-2, 0.4)$ and $\boldsymbol{\beta}_{\mathbf{\Phi}}^* = (2, -4)$. The remaining parameters are randomized while $\mathbf{\Gamma}(\tau, b = 1)$ is varied.

The results illustrate that if the degree of validity for some instruments can be bounded, `budgetIV` may be insensitive to weak constraints on other instruments. By contrast, when used for partial identification, the convex relaxations suffer from these bottlenecks of limited background knowledge, and the ATE bounds grow linearly as a result.

Budget constraints highlight possible mechanisms. Fig. 4 shows bounds on the ATE for feasible values of $\boldsymbol{U}$ with a quadratic ground truth $\Phi^*(X)$ under the presence of (A2) and (A3) violations. We study a collection of violations in Appx. D.

Feasible sets returned by `budgetIV` represent a union of convex bounds, each of which corresponds to a unique causal hypothesis that cannot be determined by the data or budget constraints alone (see Fig. 4). This drastically reduces the search space of possible causal mechanisms for the IV candidates. If one or more of these solutions can be ruled out by expert knowledge—e.g., if a monotonicity assumption is justified (Angrist et al.,
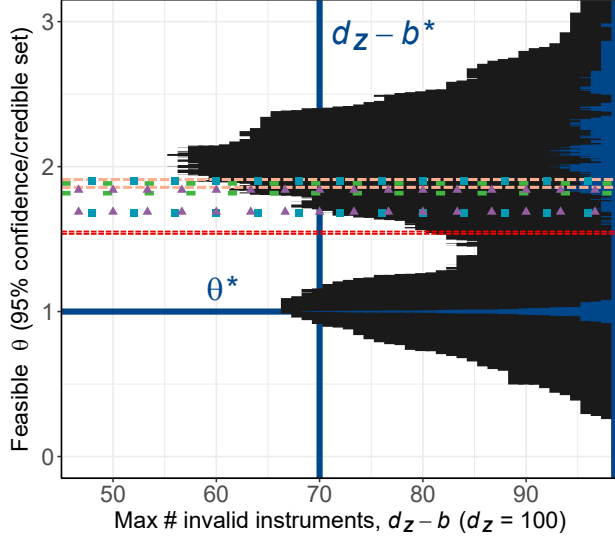
Method	Confidence/Credible Set	$\ni \theta^*$
MR-Egger	$[1.680, 1.898]$	✗
MR-Median	$[1.817, 1.884]$	✗
MBE	$[1.857, 1.911]$	✗
IVW	$[1.688, 1.839]$	✗
MASSIVE	$[1.536, 1.552]$	✗
`budgetIV`	$[0.914, 1.160] \cup [1.705, 2.403]$	✓
Oracle	$[0.997, 1.003]$	✓

Figure 5: `budgetIV` **captures the true causal effect when most candidates are invalid IVs.** Results from a simulation study with $d_{\mathbf{Z}} = 100$ candidate IVs, 70 of which violate (A3), benchmarking (Black) 95% coverage of the feasible set according to `budgetIV_scalar`, where the budget constraints $\mathbf{\Gamma}(\tau = 0, b)$ are varied along the x-axis. (Blue) The optimal solution set relative to the constraint $\mathbf{\Gamma}(\tau = 0.001, b)$ (for visibility) captures the true causal effect if and only if the choice of b doesn't exclude the ground truth $\boldsymbol{\gamma}_{\boldsymbol{g}}^*$. (Others) Confidence intervals for benchmark methods produce do not include θ^*.

1996)—then the method can be rerun with added constraints, pruning the search space still further. In this way, `budgetIV` can help practitioners evaluate causal systems in a dynamic and principled manner, aiding in hypothesis generation and experimental design.

In Fig. 4 the ground truth corresponds to the only plausible ATE with positive convexity. With additional expert knowledge, one might infer that Z_1, Z_2, and Z_3 are the true set of valid instruments.

Consistent inference under relaxed assumptions. Fig. 5 shows a benchmarking experiment with $d_{\mathbf{Z}} = 100$ candidate IVs, 30 of which are valid and the remaining 70 violate (A3). The data is simulated via a linear

with a scalar exposure X and multivariate Gaussian $(\epsilon_x, \epsilon_y, \mathbf{Z} := \boldsymbol{\epsilon_z})$. Linearity, (A3) violation, scalar X and Gaussian exogenous variables reflect the modeling assumptions in the benchmarking methods. We apply `budgetIV_scalar` for computational efficiency.

We apply a series of budget constraints $\mathbf{\Gamma}(\tau = 0, b)$ where b takes each value from 1 to 100 ($b = 0$ corresponds to no constraint). We calculate $\boldsymbol{\delta\beta_y}$ from a 95% confidence set using $100,000$ samples, which is a typical GWAS sample size.

The resulting confidence sets decompose into three kinds: ($d_{\mathbf{Z}} - b < 56$) empty, falsifying the budget constraints; ($56 \leq d_{\mathbf{Z}} - b < 67$) nonempty but excluding θ^*; and ($d_{\mathbf{Z}} - b \geq 67$) containing θ^*. Whenever $b^* < b$ (so that $\boldsymbol{\gamma}_{\boldsymbol{g}}^* \in \mathbf{\Gamma}(0, b)$) our confidence set captures θ^*. Similarly, the solution set, calculated using $\boldsymbol{\beta_y^*}$ and $\boldsymbol{\beta_\Phi^*}$, is shown to contain θ^* if and only if $b < b^*$.

The methods we benchmark against give biased results. Inverse variance weighting (IVW) is a classical method that relies on all candidates being valid IVs. MR-median (Bowden et al., 2016a) assumes $\mathrm{median}(\gamma_{g_i}^*)_{i \in [d_{\mathbf{Z}}]} = 0$. MR-Egger assumes candidates are invalid through independent mechanisms (e.g., not affecting Y via shared mediators or being confounded with Y via shared confounders). These assumptions do not hold in our simulation study so we expect the results to be biased. Intriguingly, the stated assumptions for the Bayesian approach MASSIVE (Bucur et al., 2020) and the mode based estimate MBE (Hartwig et al., 2017) hold in our experiment.

The MASSIVE estimator assumes an L_0-norm constraint $\mathbf{\Gamma}(\tau = 0, b)$ with a tuneable value of b, which we set to b^*. We have shown θ^* is only partially identified under such constraints. MASSIVE applies Bayesian model averaging over plausible sets of valid IVs. The resulting posterior distribution sits between the two disjoint confidence intervals returned by `budgetIV_scalar`, which themselves correspond to different sets of valid IVs.

MBE assumes $\mathrm{mode}(\gamma_{g_i}^*)_{i \in [d_{\mathbf{Z}}]} = 0$. Fig. 5 shows that this assumption holds because the optimal solution set peaks at $\theta = \theta^* := 1$. They estimate the corresponding modal causal effect θ using the summary statistics $\hat{\boldsymbol{\beta}}_{\boldsymbol{y}}$, $\hat{\boldsymbol{\beta}}_\Phi$, standard errors and the bandwidth selection rule of Bickel and Levina (2008). Performing statistical inference on an estimator of the mode is not straightforward (Genovese et al., 2015). Hartwig et al. (2017) base their confidence intervals a normal approximation bootstrap. However, mode estimators generally do not converge to a normal distribution, which may explain the underestimated uncertainty. Indeed Fig. 7 in Appx. D.3 shows an expanded grid of experiments under which the estimates from MBE are highly variable.

6 RELATED WORK

There is a substantial literature on partial identification for IV models with restricted outcome domains. Early work includes seminal papers by Manski (1990) and Balke and Pearl (1997). Later work considered bounded violations of the IV assumptions (Ramsahai, 2012; Silva and Evans, 2016; Jiang and Kocaoglu, 2024). Others have proposed generic methods for bounding counterfactual probabilities in discrete settings—either via polynomial programs (Duarte et al., 2023; Michael C. Sachs and Gabriel, 2023) or Markov chain Monte Carlo (Zhang et al., 2022)—with applications to IV models.

In the continuous setting, partial identification for pseudo-IV models has typically been formulated with respect to some convex relaxation of either (A2) or (A3), in both linear (Conley et al., 2012; Watson et al., 2024) and nonlinear SEMs (Newey and Powell, 2003; Gunsilius, 2019). In recent years, several authors have described more generic solutions based on stochastic gradient descent (Kilbertus et al., 2020; Hu et al., 2021; Padh et al., 2023). Unlike the linear programming approach of budgetIV, these methods are not guaranteed to converge on global optima.

In the MR literature, various methods are designed to try to handle linkage disequilibrium and/or pleiotropy. One strategy is to include a large number of candidate instruments and assume that biases will tend to cancel out in the limit (Kolesár et al., 2015; Bowden et al., 2015). Others take a feature selection approach in which Z's may be rejected on the basis of statistical tests (Chu et al., 2001; Kang et al., 2020) or L_1-penalized regression (Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019; Xue et al., 2023). Alternatively, instruments may be pooled into a single feature (Kuang et al., 2020). An especially popular choice is the modal validity assumption (Hartwig et al., 2017; Hartford et al., 2021), which asserts that the most common causal effect estimate is consistent. The goal in these works is point identification, which may be unrealistic if underlying assumptions fail.

Bayesian methods for causal inference in IV models are well-established. Priors can be used to encode uncertainty with regard to latent parameters that track either (A2) (Shapland et al., 2019) or (A3) violations (Bucur et al., 2020; Gkatzionis et al., 2021).

In recent work, Vancak and Sjölander (2023) define a single sensitivity parameter for violation of (A2) and/or (A3). We extend this approach by using a latent statistic of a nonparametric function. Star-domain restrictions for partial identification have also been proposed (Molinari, 2008), though not as a guiding principle.

For a summary of the constraints and affordances of various relaxed IV methods, see Appx. F, Table 1.

7 DISCUSSION

The budgetIV optimization problem is NP-hard for any $d_\Phi > 1$. Therefore, with multivariate Φ, the method becomes impractical for large d_Z. Approximate solutions that rely on grid search may be preferable in such cases. Many MR studies are built on just one or a handful or genetic variants, so a cap on d_Z may not be overly restrictive in such settings.

The additive separability of causal effects in pseudo-IV models has been assumed by various authors (Newey and Powell, 2003; Saengkyongam et al., 2022; Christiansen et al., 2021). We have considered the special case of a homogeneous treatment effect, though our method can be further generalized if Φ is promoted to known functions of X, Z and/or ϵ_z. Extending our approach to compute feasible sets of *conditional* average treatment effects may result in more informative outputs (Cai et al., 2007; Hartford et al., 2021; Levis et al., 2023).

We rely on the assumption $d_\Phi \leq d_Z$ and assume our choice of Φ is sufficient to describe the ground truth causal effect of X on Y exactly. Future work could could investigate error arising from misspecification of Φ or incorporate thresholds on this kind of error into the background assumptions. In particular, basis expansions that have been truncated to satisfy the bound on d_Φ may be of interest.

While we have presented our method as inferring the ATE from the distribution of instrument validity, the correspondence can be thought of as bidirectional. One promising direction for future work is to invert such methods for the purpose of instrument discovery (Silva and Shimizu, 2017). Another direction could be to use background knowledge about the functional form of the ATE (when $d_\Phi > 1$) to restrict the solution set.

There are several other extensions to budgetIV that could be of interest. Sharper partial identification may be achieved in models with restricted outcome domains. The finite-sample properties of our method might also be improved. We used a Bonferroni correction to construct the confidence set over β_y. Since this is conservative, a future direction may involve adaptively selecting the confidence thresholds for each $(\beta_y)_i$ to minimize the width of the bounds on the ATE. On the other hand, coverage under finite sample uncertainty without the NOME assumption remains an open problem.

References

Ailer, E., Hartford, J., and Kilbertus, N. (2023). Sequential underspecified instrument selection for cause-effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 408–420.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.*, 91(434):444–455.

Angrist, J. D. and Pischke, J. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

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

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Stat.*, 36(1):199 – 227.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.*, 44(2):512–525.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.*, 40(4):304–314.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.*, 45(6):1961–1974.

Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., and Davey Smith, G. (2018). Improving the accuracy of two-sample summary-data mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.*, 48(3):728–742.

Bowden, R. J. and Turkington, D. A. (1984). *Instrumental variables*. Cambridge University Press, Cambridge.

Bucur, I. G., Claassen, T., and Heskes, T. (2020). MASSIVE: Tractable and robust Bayesian learning of many-dimensional instrumental variable models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 1049–1058.

Cai, Z., Kuroki, M., and Sato, T. (2007). Nonparametric bounds on treatment effects with noncompliance by covariate adjustment. *Stat. Med.*, 26(16):3188–3204.

Christiansen, R., Pfister, N., Jakobsen, M., Gnecco, N., and Peters, J. (2021). A causal framework for distribution generalization. *IEEE transactions on pattern analysis and machine intelligence.*

Chu, T., Scheines, R., and Spirtes, P. L. (2001). Semi-instrumental variables: A test for instrument admissibility. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 83–90.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Rev. Econ. Stat.*, 94(1):260–272.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. (2023). An automated approach to causal inference in discrete settings. *J. Am. Stat. Assoc.*, pages 1–16.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2015). Non-parametric inference for density modes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1):99–126.

Gkatzionis, A., Burgess, S., Conti, D. V., and Newcombe, P. J. (2021). Bayesian variable selection with a pleiotropic loss function in Mendelian randomization. *Stat. Med.*

Gunsilius, F. (2019). A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv:1910.09502*.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. R. Stat. Soc. Series B Stat. Methodol.*, 80(4):793–815.

Hartford, J. S., Veitch, V., Sridhar, D., and Leyton-Brown, K. (2021). Valid causal inference with (some) invalid instruments. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4096–4106.

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.*, 46(6):1985–1998.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proc. Natl. Acad. Sci.*, 96(8):4730–4734.

Holland, P. W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.*, 81(396):945–960.

Hu, Y., Wu, Y., Zhang, L., and Wu, X. (2021). A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Jiang, Z. and Kocaoglu, M. (2024). Conditional common entropy for instrumental variable testing and partial identification. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21824–21843. PMLR.

Kang, H., Lee, Y., Cai, T. T., and Small, D. S. (2020). Two robust tools for inference about causal effects with invalid instruments. *Biometrics*.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). *J. Am. Stat. Assoc.*, 111(513):132–144.

Kilbertus, N., Kusner, M. J., and Silva, R. (2020). A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econ. Stat.*, 33(4):474–484.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica*, 17(2):125–144.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Re, C. (2020). Ivy: Instrumental variable synthesis for causal inference. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, page 398–410.

Levis, A. W., Bonvini, M., Zeng, Z., Keele, L., and Kennedy, E. H. (2023). Covariate-assisted bounds on causal effects with instrumental variables. *arXiv* preprint, 2301.12106.

Malloy, M. L., Tripathy, A., and Nowak, R. D. (2021). Optimal confidence regions for the multinomial parameter. *arXiv* preprint, 2002.01044.

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

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer, New York.

Michael C. Sachs, Gustav Jonzon, A. S. and Gabriel, E. E. (2023). A general method for deriving tight symbolic bounds on causal effects. *J. Comput. Graph. Stat.*, 32(2):567–576.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Nevo, A. and Rosen, A. M. (2012). Identification with imperfect instruments. 94(3):659–671.

Newey, W. and Powell, J. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. (2023). Stochastic causal programming for bounding treatment effects. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning*, pages 142–176.

Pearl, J. (2009). *Causality*. Cambridge University Press, New York.

Pfister, N. and Peters, J. (2022). Identifiability of sparse causal effects using instrumental variables. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, pages 1613–1622.

Pierce, B. L. and Burgess, S. (2013). Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184.

Ramsahai, R. R. (2012). Causal bounds and observable constraints for non-deterministic models. *J. Mach. Learn. Res.*, 13(29):829–848.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66.

Saengkyongam, S., Henckel, L., Pfister, N., and Peters, J. (2022). Exploiting independent instruments: Identification and distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18935–18958.

Shapland, C. Y., Thompson, J. R., and Sheehan, N. A. (2019). A Bayesian approach to Mendelian randomisation with dependent instruments. *Stat. Med.*, 38(6):985–1001.

Silva, R. and Evans, R. (2016). Causal inference through a witness protection program. *J. Mach. Learn. Res.*, 17(56):1–53.

Silva, R. and Shimizu, S. (2017). Learning instrumental variables with structural and non-Gaussianity assumptions. *J. Mach. Learn. Res.*, 18(120).

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vancak, V. and Sjölander, A. (2023). Sensitivity analysis of g-estimators to invalid instrumental variables. *Statistics in Medicine*, 42(23):4257–4281.

Watson, D. S., Penn, J., Gunderson, L. M., Bravo-Hermsdorff, G., Mastouri, A., and Silva, R. (2024). Bounding causal effects with leaky instruments. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *J. Am. Stat. Assoc.*, 114(527):1339–1350.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan, New York.

Xue, H., Shen, X., and Pan, W. (2023). Causal inference in transcriptome-wide association studies with invalid instruments and GWAS summary data. *Journal of the American Statistical Association*, 118(543):1525–1537.

Zhang, J., Tian, J., and Bareinboim, E. (2022). Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 26548–26558.

A   PROOFS

A.1   Proof of Thm. 1

This is a special case of a standard result from the classical IV literature, which goes back to Koopmans (1949): $\boldsymbol{\theta}$ has a unique solution under the assumptions (B1*) and (B2*) because $\boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi} = \mathbf{0}$ is a complete system of simultaneous equations. With finite samples, it is common to estimate $\boldsymbol{\theta}$ solving this equation via two-stage least squares (2SLS) or (for $d_\Phi = 1$) inverse variance weighting (IVW). The 2SLS estimator is given by:

$$\boldsymbol{\theta}_{(2\text{SLS})} := (\boldsymbol{\beta_\Phi} \cdot \Sigma_{\boldsymbol{ZZ}}^{-1} \cdot \boldsymbol{\beta_\Phi}^\top)^{-1} \cdot \boldsymbol{\beta_\Phi} \cdot \Sigma_{\boldsymbol{ZZ}}^{-1} \cdot \boldsymbol{\beta_y},$$

where $\Sigma_{\boldsymbol{ZZ}}$ denotes the $d_{\boldsymbol{Z}} \times d_{\boldsymbol{Z}}$ covariance matrix between the instruments. For details, see Wright (1928); Bowden and Turkington (1984); Angrist and Pischke (2009).

A.2   Proof of Thm. 2

Let the structural Eqs. (1) to (3) hold for some ground truth functions $f_{\boldsymbol{z}}^*$, $f_{\boldsymbol{x}}^*$, $\boldsymbol{\Phi}^*$, g_y^*, and causal parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{d_\Phi}$.

Soundness   We prove soundness of the feasible map $\mathcal{T}$ for which:

$$\mathcal{T}\left(c = \mathbb{I}[\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}], s = (\boldsymbol{\beta_\Phi}, \boldsymbol{\beta_y})\right) = \{\boldsymbol{\theta} \in \mathbb{R}^{d_\Phi} : h(\boldsymbol{\theta}) \in \boldsymbol{\Gamma}\},$$

where $h(\boldsymbol{\theta}) = \boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi}$. For ease of notation, we use the shorthand $\mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi}, \boldsymbol{\beta_y})$ hereinafter.

It follows immediately from Eq. (3) and the left-linearity of the covariance operator that:

$$\boldsymbol{\gamma_g^*} := \text{Cov}(g_y(\boldsymbol{Z}, \boldsymbol{\epsilon_y}), \boldsymbol{Z}) = \text{Cov}(Y, \boldsymbol{Z}) - \text{Cov}(\boldsymbol{\theta^*} \cdot \boldsymbol{\Phi}(\boldsymbol{X}), \boldsymbol{Z})$$
$$= \boldsymbol{\beta_y^*} - \boldsymbol{\theta^*} \cdot \boldsymbol{\beta_\Phi^*}.$$

Therefore, $\boldsymbol{\theta}^* \in \mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi^*}, \boldsymbol{\beta_y^*})$ whenever $\boldsymbol{\gamma_g^*} = (\boldsymbol{\beta_y^*} - \boldsymbol{\theta^*} \cdot \boldsymbol{\beta_\Phi^*}) \in \boldsymbol{\Gamma}$ as required.

Completeness   To prove completeness of $\mathcal{T}$ we have to show that any $\boldsymbol{\theta} \in \mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi}, \boldsymbol{\beta_y})$ cannot be rejected by any statistic of the observed joint distribution $P(\boldsymbol{Z}, \boldsymbol{X}, Y)$. Notice that in specifying Eqs. (1) to (3), we have not made any a priori assumptions about the joint distribution $P_\epsilon(\boldsymbol{\epsilon_z}, \boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y})$ or the function classes to which $f_{\boldsymbol{z}}^*$, $f_{\boldsymbol{x}}^*$, g_y^* belong. In Appx. B.1 we extend this proof to find conditions under which $\mathcal{T}$ remains complete when further structural assumptions are made.

Consider any $\boldsymbol{\theta}^\dagger \in \mathbb{R}^{d_\Phi}$. Then the following holds:

(Z) There exists at least one function $f_{\boldsymbol{z}}^\dagger$ with the following property. Given any $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either: (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$ for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ and $y \in \Omega_Y$; or (ii) there exists at least one value $\boldsymbol{\epsilon_z^\dagger}$ that solves the equation:

$$f_{\boldsymbol{z}}^\dagger(\boldsymbol{\epsilon_z^\dagger}) = \boldsymbol{z}.$$

Note that the ground truth $f_{\boldsymbol{z}}^*$ is one valid choice of $f_{\boldsymbol{z}}^\dagger$.

We can, therefore, define a function $\boldsymbol{\epsilon_z^\dagger}(\boldsymbol{z})$ that satisfies the above equation for all $\boldsymbol{z}$ for which case (i) is false. We have not demanded either $f_{\boldsymbol{z}}^\dagger$ or $\boldsymbol{\epsilon_z^\dagger}(\boldsymbol{z})$ to be unique.

(X) Likewise, there exists at least one function $f_{\boldsymbol{x}}^\dagger$ with the following property. Given any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}, \boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either: (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$ for all $y \in \Omega_Y$, or (ii) there exists at least one value $\boldsymbol{\epsilon_x^\dagger}$ that solves the equation:

$$f_{\boldsymbol{x}}^\dagger(\boldsymbol{z}, \boldsymbol{\epsilon_x^\dagger}) = \boldsymbol{x},$$

and $f_{\boldsymbol{x}}^*$ is one valid choice of $f_{\boldsymbol{x}}^\dagger$. We can, therefore, define at least one function $\boldsymbol{\epsilon_x^\dagger}(\boldsymbol{z}, \boldsymbol{x})$ that satisfies the above equation for all $\boldsymbol{x}$ and $\boldsymbol{z}$ for which case (i) is false.

(Y) Likewise, for any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}, y \in \Omega_Y, \boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$, or (ii) for any $\boldsymbol{\theta}^\dagger \in \mathbb{R}^{d_\Phi}$ there exists at least function $g_y^\dagger$ and one value $\boldsymbol{\epsilon_y^\dagger}$ that solves the equation:

$$g_y^\dagger(\boldsymbol{z}, \boldsymbol{\epsilon_y^\dagger}) = y - \boldsymbol{\theta}^\dagger \cdot \boldsymbol{\Phi}(\boldsymbol{x}).$$

For instance we could have $g_y(\boldsymbol{z}, \boldsymbol{\epsilon_y}) = \boldsymbol{\epsilon_y}$ and $\boldsymbol{\epsilon_y^\dagger} := \boldsymbol{y} - \boldsymbol{\theta^\dagger} \cdot \boldsymbol{\Phi}(\boldsymbol{x})$ is one such choice. We can therefore define a function $\boldsymbol{\epsilon_y^\dagger}(\boldsymbol{z}, \boldsymbol{x}, y; \boldsymbol{\theta^\dagger})$ for each $\boldsymbol{\theta^\dagger} \in \mathbb{R}^{d_\Phi}$ that solves the above equation for any $\boldsymbol{x}$, y and $\boldsymbol{z}$.

The full joint distribution that generates $P(\boldsymbol{X}, Y, \boldsymbol{Z})$ from the structural equations can be factorized as follows:

$$P(\boldsymbol{X}, Y, \boldsymbol{Z}, \boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) = P(\boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y} \mid \boldsymbol{X}, Y, \boldsymbol{Z}) \, P(\boldsymbol{X}, Y, \boldsymbol{Z}).$$

We can therefore define (for any $\boldsymbol{\theta^\dagger}$) at least one joint distribution consistent with the structural assumptions, $\boldsymbol{\theta^\dagger}$, and the observed joint distribution $P(\boldsymbol{X}, Y, \boldsymbol{Z})$:

$$P^\dagger(\boldsymbol{X}, Y, \boldsymbol{Z}, \boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) := D^\dagger(\boldsymbol{\epsilon_x} \mid \boldsymbol{X}, \boldsymbol{Z}) \, \delta(\boldsymbol{\epsilon_y} - \boldsymbol{\epsilon_y^\dagger}(\boldsymbol{X}, Y, \boldsymbol{Z}; \boldsymbol{\theta})) \, P(\boldsymbol{X}, Y, \boldsymbol{Z}),$$

where:

$$D^\dagger(\boldsymbol{\epsilon_x} \mid \boldsymbol{x}, \boldsymbol{z}) = \begin{cases} \delta(\boldsymbol{\epsilon_x} - \boldsymbol{\epsilon_x^\dagger}(\boldsymbol{x}, \boldsymbol{z})) & \exists y \in \Omega_Y : P(\boldsymbol{x}, y, \boldsymbol{z}) \neq 0 \\ 0 & \text{otherwise}, \end{cases}$$

and δ's are Dirac delta measures over the domains of their arguments.

Therefore, we cannot reject any $\boldsymbol{\theta} \in \mathbb{R}^{d_\Phi}$ based on the structural equations and the ground truth $P(\boldsymbol{X}, Y, \boldsymbol{Z})$ alone, since there exists some joint distribution consistent with both.

Assume the validity of a background constraint $\boldsymbol{\Gamma}$, so that $\text{Cov}(\boldsymbol{Z}, g_y) \equiv (\boldsymbol{\beta_y^*} - \boldsymbol{\theta^*} \cdot \boldsymbol{\beta_\Phi^*}) \in \boldsymbol{\Gamma}$. We can only use $\boldsymbol{\Gamma}$ to exclude $\boldsymbol{\theta}$ for which $(\boldsymbol{\beta_y^*} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi^*}) \notin \boldsymbol{\Gamma}$. Therefore, it follows immediately that the feasible map $\mathcal{T}$ for which $\mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi^*}, \boldsymbol{\beta_y^*}) = \{\boldsymbol{\theta} : (\boldsymbol{\beta_y^*} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi^*}) \in \boldsymbol{\Gamma}\}$ is complete.

Minimality We have proven that any and all $\boldsymbol{\theta} \in \mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi}, \boldsymbol{\beta_y})$ satisfy $(\boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi}) \in \boldsymbol{\Gamma}$. Since $\boldsymbol{\Gamma}$, $\boldsymbol{\beta_\Phi}$ and $\boldsymbol{\beta_y}$ can be varied arbitrarily and independently of each other, it is clear the $d_{\boldsymbol{Z}} d_\Phi + d_{\boldsymbol{Z}}$ independent, real parameters required to specify $\boldsymbol{\beta_\Phi}$ and $\boldsymbol{\beta_y}$ are needed to specify the feasible map $\mathcal{T}$. This completes the proof of optimality.

A.3 Proof of Thm. 3

We prove a slight generalization of the theorem that accounts for violation of (B1*). As we will see, the the tight bound depends on a quantity B that equals 0 when (B1*) is satisfied.

We can account for violation of (B1*) by defining a reduced problem

We extend the theorem as stated by allowing for (B1*) violation. By simple linear algebra, the left and right null spaces and accounting for (B1*) violation can be identified polynomial time.

If $\boldsymbol{\beta_\Phi}$ has a nonempty left null space, then each point $h(\boldsymbol{\theta}) \in h := h[\boldsymbol{\theta}]$ (i.e., the affine space that is the image of the affine map $h(\boldsymbol{\theta})$) corresponds to a continuum of possible $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\Phi}$, where Θ is unbounded. This means any nonempty feasible set $\mathcal{T}_{L_0}$ will also be unbounded, and depending on the choice of $\boldsymbol{\Phi}$, the ATE may be vacuous. We choose to ignore this kind of violation of (B1*), which corresponds to $\boldsymbol{\beta_\Phi}$ not being full rank.

We decompose the set of covariance-irrelevant candidate instruments $I = \{i \in [d_{\boldsymbol{Z}}] : (\boldsymbol{\beta_\Phi})_i = \boldsymbol{0}\}$ into the following subsets:

$$I_{=0} := \{i \in [d_{\boldsymbol{Z}}] : (\boldsymbol{\beta_\Phi})_i = \boldsymbol{0} \wedge (\beta_y)_i = 0\},$$
$$I_{\neq 0} := \{i \in [d_{\boldsymbol{Z}}] : (\boldsymbol{\beta_\Phi})_i = \boldsymbol{0} \wedge (\beta_y)_i \neq 0\},$$

where the former corresponds to irrelevant candidate instruments that are uncorrelated with the outcome while the latter corresponds to those correlated with the outcome. Any $\boldsymbol{\gamma_g} \in h := h[\boldsymbol{\theta}]$ will satisfy $(\gamma_g)_{i \in I_{=0}} = 0$ and $(\gamma_g)_{i \in I_{\neq 0}} \neq 0$. Therefore, we can count whether these components are always or never 0 irrespective of $\boldsymbol{\theta}$.

This motivates the following definitions:

$$D_{\boldsymbol{Z}} := d_{\boldsymbol{Z}} - |I_{=0}| - |I_{\neq 0}|,$$
$$B := b - |I_{\neq 0}|,$$

through which we define the reduced problem of finding $\boldsymbol{\theta}$ for which:

$$\|H(\boldsymbol{\theta})\|_0 \leq B,$$

where:

$$H(\boldsymbol{\theta}) = \sum_{j \in [d_{\boldsymbol{Z}}] \setminus I} \big((\beta_y)_i - \boldsymbol{\theta} \cdot (\boldsymbol{\beta_\Phi})_i \big) \boldsymbol{e}_i.$$

The solutions for $\boldsymbol{\theta}$ are exactly the same solutions as those to $\|h(\boldsymbol{\theta})\|_0 \leq b$. Notice that $H := H[\boldsymbol{\theta}]$ is a $d_{\boldsymbol{\Phi}}$-dimensional affine subspace of $\mathbb{R}^{D_{\boldsymbol{Z}}}$ (we can ignore $i \in I$). Defining $J = [d_{\boldsymbol{Z}}] \setminus I$, we see that none of the basis vectors $\boldsymbol{e}_j$ for $j \in J$ are orthogonal to H.

Notice that $D_{\boldsymbol{Z}} = d_{\boldsymbol{Z}}$, $B = b$ and $H(\boldsymbol{\theta}) = h(\boldsymbol{\theta})$ iff (B1*) holds.

We begin with $d_{\Phi} = 1$

If $d_{\Phi} = 1$ then H represents a line embedded in the $D_{\boldsymbol{Z}}$-dimensional Euclidean space. Consider a point $\boldsymbol{\gamma} := H(\theta) \in H$ for some $\theta \in \mathbb{R}$. This point satisfies $\|\boldsymbol{\gamma}\|_0 \leq B$ iff there are at least $D_{\boldsymbol{Z}} - B$ many values of $j \in J$ for which $\gamma_j = 0$.

Since there are no $j \in J$ for which H is orthogonal to J, we know there is exactly one solution $\theta_{(j)}$ to the equation $(\beta_y)_j - \theta(\beta_\Phi)_j = 0$. Thus the equivalence relation $\sim$ defined by:

$$\forall i, j \in J : i \sim j \iff \theta_{(i)} = \theta_{(j)},$$

forms a partition over $[J]$. Each equivalence class $\langle j \rangle$ of this partition must have cardinality at least $D_{\boldsymbol{Z}} - B$ for $\theta_{(j)}$ to be in the feasible set. Since $|J| = D_{\boldsymbol{Z}}$, the number of unique $\theta_{(j)}$ solving the constraint, n, is bounded above by:

$$\left\lfloor \frac{D_{\boldsymbol{Z}}}{D_{\boldsymbol{Z}} - B} \right\rfloor. \tag{5}$$

This proves Cor. 3.2, including the special case of point identification when $B > D_{\boldsymbol{Z}}/2$ first shown by Kang et al. (2016).

For $d_{\boldsymbol{\Phi}} > 1$ we enumerate the plausible unique points satisfying the budget constraints

Consider demanding a single instrument is valid In general, H corresponds to a $d_{\boldsymbol{\Phi}}$-dimensional affine space, not orthogonal to $\boldsymbol{e}_j$ for any $j \in J$. For some choice of $j \in J$, H will intersect the $(D_{\boldsymbol{Z}} - 1)$-dimensional affine space $\{\boldsymbol{\gamma} \in \mathbb{R}^{D_{\boldsymbol{Z}}} : \gamma_j = 0\}$ to form $H_{\setminus j} := \{\boldsymbol{\gamma}' \in H : \gamma_j' = 0\}$. $H_{\setminus j}$ is necessarily a $(d_{\boldsymbol{\Phi}} - 1)$-dimensional affine space by the rule for the dimension of the intersection of two affine spaces and the fact that $\boldsymbol{e}_j \not\perp H$ (where $\perp$ means orthogonal). Notice that $\boldsymbol{e}_j \perp H_{\setminus j}$.

It is possible that $H_{\setminus j} = H_{\setminus k}$ for some $k \in J \setminus j$, which motivates introducing another equivalence relation $\overset{d_{\boldsymbol{\Phi}}}{\sim}$:

$$\forall j, k \in J : j \overset{d_{\boldsymbol{\Phi}}}{\sim} k \iff H_{\setminus j} = H_{\setminus k},$$

which induces a partition over J. The equivalence classes $\langle j \rangle_{d_{\boldsymbol{\Phi}}}$ represent a single $(d_{\boldsymbol{\Phi}} - 1)$-dimensional affine space that consists of points $\boldsymbol{\gamma}$ that simultaneously satisfy the constraints $\gamma_k = 0$ for all k in the equivalence class. In other terms:

$$H_{\setminus j} = H_{\setminus \langle j \rangle_{d_{\boldsymbol{\Phi}}}} := \{\boldsymbol{\gamma} \in H : \forall k \in \langle j \rangle_{d_{\boldsymbol{\Phi}}} (\gamma_k = 0)\}.$$

The number of unique equivalence classes with respect to $\overset{d_{\boldsymbol{\Phi}}}{\sim}$, $n_{d_{\boldsymbol{\Phi}}}$, is anywhere between 1 and $D_{\boldsymbol{Z}}$. However, if any of the equivalence classes has cardinality at least $D_{\boldsymbol{Z}} - B$, then there is at least one $(d_{\boldsymbol{\Phi}} - 1)$-dimensional solution to the L_0-norm constraint.

What if we demand more instruments are valid? We could equally ask how many sets of constraints $(\gamma_q)_{q \in Q} = 0$ for $q \in Q \subseteq J$ lead to unique $(d_{\boldsymbol{\Phi}} - 2)$-dimensional affine spaces $H_{\setminus Q} := \{\boldsymbol{\gamma} \in H : \forall k \in Q (\gamma_k = 0)\}$. The Q such that $H_{\setminus Q}$ is a $(d_{\boldsymbol{\Phi}} - 2)$-dimensional affine space is the union of at least two equivalence classes $\langle j \rangle_{d_{\boldsymbol{\Phi}}}$ and $\langle k \rangle_{d_{\boldsymbol{\Phi}}} \neq \langle j \rangle_{d_{\boldsymbol{\Phi}}}$. Therefore the number of unique $(d_{\boldsymbol{\Phi}} - 2)$-dimensional affine spaces $H_{\setminus Q}$, denoted $n_{d_{\boldsymbol{\Phi}} - 1}$, is anywhere between 1 and $\binom{n_{d_{\boldsymbol{\Phi}}}}{2}$.

Likewise, the number of sets Q generating unique 1-dimensional affine spaces (lines) $H_{\backslash Q}$, denoted n_1, is anywhere between 1 and $\binom{n_{d_\Phi}}{d_\Phi - 1}$. This implies the upper bound:

$$n_1 \leq \binom{D_Z}{d_\Phi - 1} = \frac{D_Z!}{(d_\Phi - 1)!(D_Z - d_\Phi + 1)!},$$

where this upper bound corresponds to all such $Q \subset J$ having length $d_\Phi - 1$, which is the minimum length Q can take.

Bounding the number of unique points for a given line   Consider some Q for which $H_{\backslash Q} := \{\gamma \in H : \forall k \in Q(\gamma_k = 0)\}$ is a line. Then define $Q' := \{i \in J : \forall \gamma \in H_{\backslash Q}(\gamma_k = 0)\}$, which expresses every constraint satisfied along the line. We ask how many unique points along $H_{\backslash Q}$ can satisfy the L_0-norm constraint $\|H_{\backslash Q}\|_0 \leq b$. Along the entire line, there are $|Q'|$ many components for which $\gamma_k = 0$. For a point to satisfy the constraint, we must have at least $D_Z - |Q'|$ many $\gamma_k = 0$ for $k \in J \backslash Q'$. Therefore, by the same arguments that lead to Eq. (5), the number of unique points solving the constraint along H_Q, denoted p_Q, is bounded by:

$$p_Q \leq \left\lfloor \frac{D_Z - |Q'|}{D_Z - |Q'| - B} \right\rfloor.$$

Putting the pieces together   As discussed earlier, the minimum length of Q' is $d_\Phi - 1$. This value maximizes p_Q. The maximum number of unique lines n_1 also increases as the length $|Q'|$ for each line decreases. Therefore, the upper bound:

$$n \leq \frac{D_Z!}{(d_\Phi - 1)!(D_Z - d_\Phi + 1)!} \left\lfloor \frac{D_Z - d_\Phi + 1}{D_Z - d_\Phi + 1 - B} \right\rfloor,$$

which corresponds to all unique lines H_Q having $Q' = d_\Phi - 1$, is valid and tight.

A.4   Proof of Thm. 4

We begin with the assumption that the estimated statistic $\hat{\beta}_\Phi$ is exactly equal to the observable ground truth β_Φ^*. Furthermore, since elliptical confidence sets require estimating $\mathrm{Cov}\left((\hat{\beta}_y)_i, (\hat{\beta}_y)_j\right)$ for pairs $i \neq j$—estimates for these are not openly available with GWAS summary statistics—we construct box constraints using the Bonferroni correction. Our $(1 - \alpha) \times 100\%$ confidence set over β_y, consists of all β_y for which each component $(\beta_y)_i$ is in the corresponding $(1 - \alpha/d_Z) \times 100\%$ confidence interval, defined below.

Calculate the estimator $\hat{\beta}_y := \widehat{\mathrm{Cov}}(Y, Z)$ using N i.i.d. samples from $P(X, Y, Z)$. It is well known from the central limit theorem (Vaart, 1998) that if this estimator has finite marginal standard errors $\mathrm{Se}(\hat{\beta}_y)_i = \sqrt{\mathrm{Var}(\hat{\beta}_y)_i}$, then the following convergence in distribution holds as our choice of N approaches ∞:

$$\frac{\sqrt{N}\left((\hat{\beta}_y)_i - (\beta_y^*)_i\right)}{\mathrm{Se}(\hat{\beta}_y)_i} \xrightarrow{d} \mathcal{N}(0, 1),$$

where β_y^* denotes the ground truth. A similar statement can be made about β_Φ under the assumption that $\mathrm{Se}(\beta_\Phi)_{ij}$ are finite for all $i \in [d_\Phi], j \in [d_Z]$. However, by the NOME assumption, we choose to neglect any variation in $\hat{\beta}_\Phi$.

We model $\mathrm{Se}(\hat{\beta}_y)_i$ with the plug in estimator $\widehat{\mathrm{Se}}(\hat{\beta}_y)_i$ and construct the required confidence intervals using quantiles of $\mathcal{N}(0, 1)$. The confidence intervals are symmetric about $(\hat{\beta}_y)_i$ and have width $2(\delta\beta_y)_i$ and we denote the resulting confidence set B_y^α.

Since $\mathcal{T}(C = \{\gamma_g \in \Gamma\}, s = \{\beta_\Phi^*, \hat{\beta}_y\})$ is a deterministic functional of $\hat{\beta}_y$, we can compute a confidence set over $\mathcal{T}$ with asymptotically valid coverage explicitly:

$$\mathcal{T}^\alpha\left(c = \mathbb{I}(\gamma_g \in \Gamma), s = \{\beta_\Phi^*, B_y^\alpha\}\right) = \left\{\theta \in \mathbb{R}^{d_\Phi} : \exists \beta_y \in B_y^\alpha\left((\beta_y - \theta \cdot \beta_\Phi^*) \in \Gamma\right)\right\}.$$

Confidence sets over a functional that are constructed in this way appear, for example, in Duarte et al. (2023) and Malloy et al. (2021).

We can rewrite $\mathcal{T}^\alpha(c, s)$ defined above as the following set:

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^{d_\Phi} : \exists \boldsymbol{\beta_y} \in B_y^\alpha, \tilde{\boldsymbol{U}} \in \Sigma_b^{(\max)} \left((\boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta}_\Phi^*) \in \tilde{\boldsymbol{\Gamma}}_{\tilde{U}} \right) \right\}.$$

For any particular $\tilde{\boldsymbol{U}} \in \Sigma_b$, define:

$$\tau_i^{\tilde{U}} = \tau_\ell \iff \tilde{U}_i = \ell.$$

Then $\boldsymbol{\theta} \in \mathcal{T}^\alpha(c, s)$, iff there exists a $\tilde{\boldsymbol{U}} \in \Sigma_b$ for which:

$$\left| (\hat{\beta}_y)_i \pm (\delta\beta_y)_i - \boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}}_X \right| \le \tau_i^{\tilde{U}} \iff \left| (\hat{\beta}_y)_i - \boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}}_X \right| \le \tau_i^{\tilde{U}} + (\delta\beta_y)_i.$$

We can therefore define $\hat{\boldsymbol{\Gamma}}_\alpha := \bigcup_{\tilde{U} \in \Sigma_b} \hat{\boldsymbol{\Gamma}}_{\tilde{U}}^\alpha$, where:

$$\hat{\boldsymbol{\Gamma}}_{\tilde{U}}^\alpha := \left\{ \boldsymbol{\gamma} \in \mathbb{R}^{d_Z} : \forall i \in [d_Z] \left(|\gamma_i| \le \tau_i^{\tilde{U}} + (\delta\beta_y)_i \right) \right\},$$

and the confidence set can be rewritten as:

$$\mathcal{T}^\alpha \left(c = \mathbb{I}[\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}], s = \{\boldsymbol{\beta}_\Phi^*, \boldsymbol{\beta_y}^{\square}\} \right) = \mathcal{T} \left(c' = \mathbb{I}[\boldsymbol{\gamma_g} \in \hat{\boldsymbol{\Gamma}}_\alpha], s = \{\boldsymbol{\beta}_\Phi^*, \hat{\boldsymbol{\beta}}_y\} \right).$$

Because $\mathcal{T}^\alpha$ includes all $\boldsymbol{\theta} \in \mathcal{T}(c, s = \{\boldsymbol{\beta}_\Phi^*, \boldsymbol{\beta}_y^*\})$ with probability at least $(1 - \alpha) \times 100\%$, we have the following guarantee:

$$P\left(\boldsymbol{\theta}^* \in \mathcal{T} \left(c', s = \{\boldsymbol{\beta}_\Phi^*, \hat{\boldsymbol{\beta}}_y\} \right) \right) \ge 1 - \alpha,$$

provided $\boldsymbol{\gamma}_g^* \in \boldsymbol{\Gamma}$ (i.e., for all $c \in \mathcal{C}^*$).

B   ADDITIONAL THEOREMS

B.1   A Condition for the Optimality of the Feasible Map under Stronger Structural Assumptions

Here we show that with stricter structural assumptions, sharper feasible maps can be obtained. In particular, we focus on the case in which the model class $\mathcal{M}'$ consists of all SCMs satisfying Eqs. (1) to (3) for some predetermined restricted function classes $\mathcal{F}_z \ni f_z$, $\mathcal{F}_x \ni f_x$ and $\mathcal{G}_y \ni g_y$.

We show that restrictions on $\mathcal{F}_z$ and $\mathcal{F}_x$ alone do not affect the sharpness of $\mathcal{T}$ and we construct necessary and sufficient conditions for restrictions over $\mathcal{G}_y$ to lead to the existence of sharper feasible maps.

The fact that $\mathcal{G}_y$ is the important restriction mirrors results in the (valid) IV literature, in which nonparameteric bounds can be put on treatment effects based solely on outcomes Y being categorical (Balke and Pearl, 1997) or bounded continuous (Manski, 1990). In fact, the following result shows that if Ω_Y is any proper subset of the real line (e.g., positive but not absolutely bounded), then a sharper feasible map exists.

Theorem 5 (Necessary and sufficient condition for optimality under stricter assumptions). Let $\mathcal{M}'$ be all SCMs consistent with Eqs. (1) to (3) under restricted function classes as described above. The feasible map $\mathcal{T}$ described in Thm. 2 remains sound against the constraint set $\mathcal{C} := \{\mathbb{I}(\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}) : \boldsymbol{\Gamma} \in \mathbb{R}^{d_Z}\}$ under any such $\mathcal{M}'$. Furthermore, the map remains complete if and only if for all $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, the section of g_y at that $\boldsymbol{z}$ is *onto* the full real line $\mathbb{R}$.

Corollary 5.1. (Restricted outcome domain) In particular, $\mathcal{T}$ is not complete if Ω_Y is a bounded subset of $\mathbb{R}$. This includes if Ω_Y is bounded and categorical, e.g., $\Omega = \{0, 1\}$.

For any $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, the "section" of g_y at $\boldsymbol{Z} = \boldsymbol{z}$ is defined as the function $g_{y|\boldsymbol{z}}$ defined over the domain $\Omega_{\boldsymbol{\epsilon_y}}$ for which $g_{y|\boldsymbol{z}}(\boldsymbol{\epsilon_y}) = g_y(\boldsymbol{z}, \boldsymbol{\epsilon_y})$. The necessary and sufficient condition for completeness is that the image of each $g_{y|\boldsymbol{z}}$, denoted $g_{y|\boldsymbol{z}}[\epsilon_y]$, is the full real line $\mathbb{R}$. Before proving the theorem, we introduce an intuitive example when this image restriction does not hold.

B.1.1 An Example in which the Map is Incomplete

Suppose $\mathcal{M}'$ requires that at some point $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, the section:

$$g_{y|\boldsymbol{z}}(\epsilon_y) := g_y(\boldsymbol{Z} = \boldsymbol{z}, \epsilon_y) = \frac{a}{1 + e^{-\epsilon_y}},$$

for some constant a. Intuitively, at this value $\boldsymbol{z}$, the proportion of $\mathrm{Var}(Y)$ due to confounding is restricted: there is a value of $\boldsymbol{z}$ for which we know the confounding between $\boldsymbol{X}$ and Y is restricted.

Then, for any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ and $y \in \Omega_Y$ for which $P(\boldsymbol{X} = \boldsymbol{x}, Y = y \mid \boldsymbol{Z} = \boldsymbol{z}) \neq 0$, we may write:

$$y = \boldsymbol{\theta} \cdot \boldsymbol{\Phi}(\boldsymbol{x}) + \frac{a}{1 + e^{\epsilon_y}}. \tag{6}$$

Assuming $\boldsymbol{\Phi}(\boldsymbol{x}) \neq 0$, the above equation imposes a restriction on $\boldsymbol{\theta}$. To see this, consider the case whereby $d_{\boldsymbol{\Phi}} = 1$ and notice that:

$$\frac{1}{1 + e^{-\epsilon_y}} \in [0, 1].$$

Then, the scalar θ is bounded by Eq. (6):

$$\theta \in \left[\frac{y}{\Phi(\boldsymbol{x})}, \frac{y}{\Phi(\boldsymbol{x})} + \frac{a}{\Phi(\boldsymbol{x})} \right].$$

Therefore, with only a single value of $\boldsymbol{z}$, a strong restriction on θ has been imposed through the structural equations and $P(\boldsymbol{X}, Y \mid \boldsymbol{Z} = \boldsymbol{z})$ alone.

B.1.2 Proof of Thm. 5

This proof mirrors that of the original optimality result Thm. 2 in Appx. A.2. We explicitly describe the effect of the restrictions $f_{\boldsymbol{z}} \in \mathcal{F}_{\boldsymbol{z}}$, $f_{\boldsymbol{x}} \in \mathcal{F}_{\boldsymbol{x}}$ and $g_y \in \mathcal{G}_y$ at each stage.

Let the structural Eqs. (1) to (3) hold for some ground truth functions $f_{\boldsymbol{z}} \in \mathcal{F}_{\boldsymbol{z}}$, $f_{\boldsymbol{x}} \in \mathcal{F}_{\boldsymbol{x}}$, $\boldsymbol{\Phi}^*$, $g_y \in \mathcal{G}_y$ and causal parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{d_{\boldsymbol{\Phi}}}$.

Soundness   The proof of soundness is not affected by the functional restrictions.

It follows immediately from Eq. (3) and the left-linearity of the covariance operator that:

$$\boldsymbol{\gamma}_g^* := \mathrm{Cov}(g_y(\boldsymbol{Z}, \boldsymbol{\epsilon_y}), \boldsymbol{Z}) = \mathrm{Cov}(Y, \boldsymbol{Z}) - \mathrm{Cov}(\boldsymbol{\theta}^* \cdot \boldsymbol{\Phi}(\boldsymbol{X}), \boldsymbol{Z})$$
$$= \boldsymbol{\beta}_{\boldsymbol{y}}^* - \boldsymbol{\theta}^* \cdot \boldsymbol{\beta}_{\boldsymbol{\Phi}}^*.$$

Therefore, $\boldsymbol{\theta}^* \in \mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta}_{\boldsymbol{\Phi}}^*, \boldsymbol{\beta}_{\boldsymbol{y}}^*)$ whenever $\boldsymbol{\gamma}_g^* = (\boldsymbol{\beta}_{\boldsymbol{y}}^* - \boldsymbol{\theta}^* \cdot \boldsymbol{\beta}_{\boldsymbol{\Phi}}^*) \in \boldsymbol{\Gamma}$ as required.

Completeness   Consider any $\boldsymbol{\theta}^\dagger \in \mathbb{R}^{d_{\boldsymbol{\Phi}}}$. Then the following holds:

(Z) There exists at least one function $f_{\boldsymbol{z}}^\dagger$ with the following property. Given any $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either: (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$ for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ and $y \in \Omega_Y$; or (ii) there exists at least one value $\boldsymbol{\epsilon}_{\boldsymbol{z}}^\dagger$ that solves the equation:

$$f_{\boldsymbol{z}}^\dagger(\boldsymbol{\epsilon}_{\boldsymbol{z}}^\dagger) = \boldsymbol{z}.$$

We know this must hold because the ground truth $f_{\boldsymbol{z}}^* \in \mathcal{F}_{\boldsymbol{z}}$ is a valid choice of $f_{\boldsymbol{z}}^\dagger$.

We can, therefore, define a function $\boldsymbol{\epsilon}_{\boldsymbol{z}}^\dagger(\boldsymbol{z})$ that satisfies the above equation for all $\boldsymbol{z}$ for which case (i) is false. We have not demanded either $f_{\boldsymbol{z}}^\dagger$ or $\boldsymbol{\epsilon}_{\boldsymbol{z}}^\dagger(\boldsymbol{z})$ to be unique.

(X) Likewise, there exists at least one function $f_{\boldsymbol{x}}^\dagger$ with the following property. Given any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}, \boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either: (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$ for all $y \in \Omega_Y$, or (ii) there exists at least one value $\boldsymbol{\epsilon}_{\boldsymbol{x}}^\dagger$ that solves the equation:

$$f_{\boldsymbol{x}}^\dagger(\boldsymbol{z}, \boldsymbol{\epsilon}_{\boldsymbol{x}}^\dagger) = \boldsymbol{x},$$

and, again, $f_{\boldsymbol{x}}^* \in \mathcal{F}_{\boldsymbol{x}}$ is a valid choice of $f_{\boldsymbol{x}}^\dagger$. We can, therefore, define at least one function $\boldsymbol{\epsilon}_{\boldsymbol{x}}^\dagger(\boldsymbol{z}, \boldsymbol{x})$ that satisfies the above equation for all $\boldsymbol{x}$ and $\boldsymbol{z}$ for which case (i) is false.

(Y) In the proof for Thm. 2, we showed that for any $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$, $y \in \Omega_Y$, $\boldsymbol{z} \in \Omega_{\boldsymbol{Z}}$, either (i) $P(\boldsymbol{x}, y, \boldsymbol{z}) = 0$, or (ii) for any $\boldsymbol{\theta}^\dagger \in \mathbb{R}^{d_\Phi}$ there exists at least one value $\boldsymbol{\epsilon}_{\boldsymbol{y}}^\dagger$ that solves the equation:

$$g_y^\dagger(\boldsymbol{z}, \boldsymbol{\epsilon}_{\boldsymbol{y}}^\dagger) = y - \boldsymbol{\theta}^\dagger \cdot \boldsymbol{\Phi}(\boldsymbol{x}).$$

However, for some choices of $\mathcal{G}_y$ this is not true. The statement can only be true if, whenever $\boldsymbol{\Phi}(\boldsymbol{x}) \neq \boldsymbol{0}$, we have $g_{y|\boldsymbol{z}}[\boldsymbol{\epsilon}_{\boldsymbol{y}}] = \mathbb{R}$. Otherwise, there is a choice of $\boldsymbol{\theta}^\dagger$ and a value $g \in \mathbb{R} \backslash g_{y|\boldsymbol{z}}[\boldsymbol{\epsilon}_{\boldsymbol{y}}]$ such that $y - \boldsymbol{\theta}^\dagger \cdot \boldsymbol{\Phi}(\boldsymbol{x})$.

We define the set of potentially problematic samples:

$$\boldsymbol{\Omega}_{\mathfrak{Y}} = \{(\boldsymbol{x}, y, \boldsymbol{z}) \in \boldsymbol{\Omega}_{\boldsymbol{x}} \times \Omega_Y \times \boldsymbol{\Omega}_{\boldsymbol{Z}} : P(\boldsymbol{X}, Y, \boldsymbol{Z}) \neq 0 \wedge \boldsymbol{\Phi}(\boldsymbol{x}) \neq \boldsymbol{0}\}, \tag{7}$$

and then the values that are "missing" in at least one section of g_y:

$$\mathfrak{G} = \bigcup_{(\boldsymbol{x}, y, \boldsymbol{z}) \in \Omega_{\mathfrak{Y}}} \left\{ \mathfrak{g} \in \mathbb{R} \backslash g_{y|\boldsymbol{z}}[\epsilon_{\boldsymbol{y}}] \right\}. \tag{8}$$

For each $\mathfrak{g} \in \mathfrak{G}$ there are a set of disallowed values for $\boldsymbol{\theta}$:

$$\Theta_{\mathfrak{g}} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{d_\Phi} : \exists (\boldsymbol{x}, y, \boldsymbol{z}) \in \Omega_{\mathfrak{Y}} \, (\boldsymbol{\theta} \cdot \boldsymbol{\Phi}(\boldsymbol{x})) = y - \mathfrak{g} \right\},$$

from which we define $\Theta_{\mathfrak{G}} := \bigcup_{\mathfrak{g} \in \mathfrak{G}} \Theta_{\mathfrak{g}}$.

We can therefore construct the sharper feasible set than $\mathcal{T} = \{\boldsymbol{\theta} : h(\boldsymbol{\theta}) \in \boldsymbol{\Gamma}\}$ by excluding every $\boldsymbol{\theta}$ which is disallowed:

$$\mathcal{T}'(c = \mathbb{I}[\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}], s' = P(\boldsymbol{X}, Y, \boldsymbol{Z})) = \mathcal{T} \backslash \Theta_{\mathfrak{G}},$$

with $\mathcal{T}' = \mathcal{T}$ iff for all $\boldsymbol{z} \in \Omega_{\boldsymbol{z}}$, $g_{y|\boldsymbol{z}}[\boldsymbol{\epsilon}_{\boldsymbol{y}}] = \mathbb{R}$ (equivalently, $\mathfrak{G} = \emptyset$).

We know the corresponding feasible map $\mathcal{T}'$ is sound because it only excludes $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{Z}}}$ if either (i) the sound feasible map $\mathcal{T}$ excludes $\boldsymbol{\theta}$, or (ii) there does not exist a $g_y \in \mathcal{G}_y$ (and thus an $m \in \mathcal{M}'$) which is consistent with $P(\boldsymbol{x}, y, \boldsymbol{z})$. We have therefore proven the necessity of the assumption that the image of each section of g_y is the full real line for sharpness of $\mathcal{T}$.

Sufficiency is simple to prove by following step (Y) in the proof of Thm. 2 in Appx. A.2 to construct a full joint distribution consistent with each $P(\boldsymbol{X}, Y, \boldsymbol{Z})$ and the SCM:

$$P^\dagger(\boldsymbol{X}, Y, \boldsymbol{Z}, \boldsymbol{\epsilon_x}, \boldsymbol{\epsilon_y}) := D^\dagger(\boldsymbol{\epsilon_x} \mid \boldsymbol{X}, \boldsymbol{Z}) \, \delta(\boldsymbol{\epsilon_y} - \boldsymbol{\epsilon}_{\boldsymbol{y}}^\dagger(\boldsymbol{X}, Y, \boldsymbol{Z}; \boldsymbol{\theta})) \, P(\boldsymbol{X}, Y, \boldsymbol{Z}).$$

For definitions of the δ's and D, visit this part of the appendix.

This concludes the proof of Thm. 5.

Minimality We note that $\mathcal{T}$ is not necessarily minimal for general $\mathcal{F}_{\boldsymbol{z}}$, $\mathcal{F}_{\boldsymbol{x}}$ and $\mathcal{G}_y$. If $\mathcal{T}$ is sharp, it may not be minimal because underlying symmetries in these function classes may constrain which $\boldsymbol{\beta_y}$, $\boldsymbol{\beta_\Phi}$ can arise, which allow one to describe the solution set $\{\boldsymbol{\theta} : (\boldsymbol{\beta_y^*} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi^*}) \in \boldsymbol{\Gamma}\}$ with fewer parameters. On the other hand, $\boldsymbol{\beta_y}$, $\boldsymbol{\beta_\Phi}$ may be insufficient for specifying $\mathcal{T}'$ when $g_{y|\boldsymbol{z}}[\boldsymbol{\epsilon_y}] \subset \mathbb{R}$.

B.2 Polytime Testability and Necessary Condition for Unidentifiability

We say the ground truth causal effect $\boldsymbol{\theta^*}$ is *unidentifiable* when the feasible set $\mathcal{T}(\boldsymbol{\Gamma}, \boldsymbol{\beta_\Phi^*}, \boldsymbol{\beta_y^*}) = \mathbb{R}^{d_\Phi}$. In plain English, this means nothing can be learned about $\boldsymbol{\theta^*}$ from the data under the current assumptions.

In this section we show that budget constraints enable an efficient test for unidentifiable causal effects even when $d_\Phi > 1$. We also show unidentifiability can only occur under budget constraints if assumption (B1*) is violated (recall the definition from Sect. 2.1).

Theorem 6 (Unidentifiability). Assume Eqs. 1, 2, 3 and budget constraints according to some $\boldsymbol{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$. Then unidentifiability of $\boldsymbol{\theta^*}$ can be decided in $\mathcal{O}(K d_{\boldsymbol{Z}} d_\Phi)$ time. Moreover, unidentifiability never occurs under (B1*).

B.2.1 Proof of Thm. 6

We prove necessary and sufficient conditions for $\mathcal{T}(\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b}), \boldsymbol{\beta_\Phi}, \boldsymbol{\beta_y}) = \mathbb{R}^{d_\Phi}$.

Defining the function $\boldsymbol{V} : \mathbb{R}^{d_Z} \mapsto [K+1]^{d_Z}$ through its components:

$$V_i(\boldsymbol{\gamma_g}) = \begin{cases} 1 & |(\gamma_g)_i| \leq \tau_1 \\ l \in \{2, 3, \ldots, K\} & \tau_{l-1} \leq |(\gamma_g)_i| \leq \tau_l \\ K+1 & |(\gamma_g)_i| > \tau_K, \end{cases}$$

we have $\boldsymbol{V}(\boldsymbol{\gamma_g^*}) = \boldsymbol{U}^*$ is the ground truth for the latent variable $\boldsymbol{U}$. We can represent $\boldsymbol{V}$ by its one-hot encoding:

$$V_{il}(\boldsymbol{\gamma_g}) = \begin{cases} 1 & V_i(\boldsymbol{\gamma_g}) \leq l \\ 0 & \text{otherwise.} \end{cases}$$

Then budget background search space is explicitly written as:

$$\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b}) := \left\{ \boldsymbol{\gamma_g} \in \mathbb{R}^{d_Z} : \forall l \in [K] \left(\sum_{i=1}^{d_Z} V_{il}(\boldsymbol{\gamma_g}) \geq b_l \right) \right\}.$$

Unidentifiability occurs iff $h(\boldsymbol{\theta}) = \boldsymbol{\beta_y} - \boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi}$ is contained within $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ for any $\boldsymbol{\theta} \in \mathbb{R}^{d_\Phi}$. Equivalently, unidentifiability occurs iff the affine space $h := h[\boldsymbol{\theta}]$ formed by the image of $h(\boldsymbol{\theta})$ is contained within $\mathbf{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$.

Consider the standard orthonormal basis $\{\boldsymbol{e}_i\}_{i=1}^{d_Z}$ for which $\boldsymbol{\gamma_g} = \sum_{i=1}^{d_Z} (\gamma_g)_i \boldsymbol{e}_i$. For any direction $\boldsymbol{e}_i$ we have that either (i) $(\beta_y)_i - (\boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi})_i = (\beta_y)_i$ is constant and h is orthogonal to $\boldsymbol{e}_i$, or (ii) $(\beta_y)_i - (\boldsymbol{\theta} \cdot \boldsymbol{\beta_\Phi})_i$ is unbounded. If (i) holds, it must be the case that $(\boldsymbol{\beta_\Phi})_i = \boldsymbol{0}$ and thus (B1*) is violated, and we say candidate instrument Z_i is covariance-irrelevant to $\boldsymbol{\Phi}(\boldsymbol{X})$.

We can define the index set for all covariance-irrelevant candidate instruments:

$$I := \{i \in [d_\Phi] : (\boldsymbol{\beta_\Phi})_i = \boldsymbol{0}\}.$$

Then, h is unbounded in all directions $[d_Z] \setminus I$ but fixed at a single value for each direction in I. Since h is affine it is also unbounded in the direction $\sum_{i \in [d_Z] \setminus I} \boldsymbol{e}_i$. Moreover, since points in h are parameterized by $\boldsymbol{\theta}$, there exists values of $\boldsymbol{\theta}$ for which $V_i(h(\boldsymbol{\theta})) = K+1$ for all $i \in [d_Z] \setminus I$.

At such a value for $\boldsymbol{\theta}$, we have $\boldsymbol{V}(h(\boldsymbol{\theta})) = \boldsymbol{V}(\boldsymbol{P})$, where $\boldsymbol{P} \in \mathbb{R}^{d_Z}$ is defined by:

$$P_i = \begin{cases} (\beta_y)_i & i \in I \\ \tau_K + 1 & \text{otherwise.} \end{cases}$$

Therefore, the necessary and sufficient condition for unidentifiability is:

$$\sum_{i=1}^{d_Z} V_{il}(\boldsymbol{P}) \geq b_l,$$

where I and $\boldsymbol{P}$ can be computed and the condition evaluated by straightforward linear algebra in $\mathcal{O}(K d_Z d_\Phi)$ time.

C ALGORITHMS

Alg. 1 depicts a combinatorial search algorithm that finds the exact feasible set of $\boldsymbol{\theta}$ and the resulting feasible set of ATE$(\boldsymbol{x}; \boldsymbol{x_0})$ subject to budget background constraints. The ATE is recovered using a grid search and program over the linear weights $\boldsymbol{\theta}$ with plug-in values of $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$.

Alg. 2 depicts a polynomial time algorithm for the case of $d_\Phi = 1$. This method utilises the single solution to the equation $A_i - B_i \theta = \tau$ for $B_i = 0$, and our ability to order such solutions along $\mathbb{R}$.

Algorithm 1 `budgetIV` general case.

1: **for** $\tilde{U} \in \Sigma_{\boldsymbol{b}}^{(\max)}$ **do**   ▷ Iterate through search space (combinatorial)
2:   **if** $h(\boldsymbol{\theta}) \in \tilde{\boldsymbol{\Gamma}}_{\tilde{U}}$ for some $\boldsymbol{\theta}$ **then**   ▷ Linear CSP with convex constraints
3:     $\check{\boldsymbol{U}}(\tilde{U}) \leftarrow \boldsymbol{0}$
4:     **for** $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$ **do**   ▷ Grid evaluation (for $d_{\boldsymbol{X}} \ll d_{\boldsymbol{\Phi}}$)
5:       Calculate $\mathrm{ATE}_{\check{U}}^{+/-}(\boldsymbol{x}; \boldsymbol{x_0})$ ▷ Linear program with convex constraints
6:         $\boldsymbol{\theta}_{\check{U}}^{+/-}(\boldsymbol{x}; \boldsymbol{x_0})$ ▷ Arguments for above LP
7:       **for** $i \in [d_{\boldsymbol{Z}}]$ **do**
8:         $\check{U}_i(\check{U}) \leftarrow \max\{\check{U}_i(\tilde{U}), U_i(\boldsymbol{\theta}^+), U_i(\boldsymbol{\theta}^-)\}$
9:       **end for**
10:     **end for**
11:   **end if**
12: **end for**
13: **return** $\left\{ \left(\mathrm{ATE}_{\check{U}}^{+/-}, \check{\boldsymbol{U}}(\tilde{U}) \right) : \tilde{U} \in \Sigma_{\boldsymbol{b}}^{(\max)} \right\}$

Algorithm 2 Polytime `budgetIV` with $d_\Phi = 1$.

1: Let $c'(\theta)$ be the indicator function for $h(\theta) \in \boldsymbol{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ and $\boldsymbol{U}(\theta)$ be the unique value of $\boldsymbol{U}$ for which $h(\theta) \in \boldsymbol{\Gamma}_{\boldsymbol{U}}(\boldsymbol{\tau}, \boldsymbol{b})$ provided $c'(\theta)$.
2: **for** $(i, j) \in [d_{\boldsymbol{Z}}] \times [K]$ **do**
3:   **if** $(\beta_\Phi)_i = 0 \wedge (\beta_y)_i \leq \tau_j$ **then**
4:     $b_j \leftarrow b_j - 1$
5:   **else if** $(\beta_\Phi)_i \neq 0$ **then**
6:     $\theta_{ij}^{\pm} \leftarrow \frac{(\beta_y)_i}{(\beta_\Phi)_i} \mp \frac{\tau_j}{(\beta_\Phi)_i}$
7:   **end if**
8: **end for**
9: **if** $\forall i \in [K](b_i \leq 0)$ **then**
10:   **return** *unidentifiable*
11: **end if**
12: $\boldsymbol{\Theta} \leftarrow \mathrm{sort}(\theta_{ij}^{\pm})$
13: $\boldsymbol{\Theta}' \leftarrow \left(\Theta_1 - 1, \frac{\Theta_2 + \Theta_1}{2}, \frac{\Theta_3 + \Theta_2}{2}, \ldots, \frac{\Theta_{|\boldsymbol{\Theta}|} + \Theta_{|\boldsymbol{\Theta}|-1}}{2}, \Theta_{|\boldsymbol{\Theta}|} + 1 \right)$
14: intervals $\leftarrow \{ \langle [\Theta_i, \Theta_{i+1}], \boldsymbol{U}(\Theta_i') \rangle : c'(\Theta_i') \}$
15: points $\leftarrow \{ \langle \Theta_i, \boldsymbol{U}(\Theta_i) \rangle : c'(\Theta_i) \wedge \neg c'(\Theta_i') \neg c'(\Theta_{i+1}') \}$
16: **if** intervals $\cup$ points $== \emptyset$ **then**
17:   **return** *infeasible*
18: **else**
19:   **return** intervals $\cup$ points
20: **end if**

D EXPERIMENTS

D.1 Linear Simulation Study

Fig. 3 is the result of a simple experiment to understand the differences between the nonconvex budget background constraints and any convex relaxation thereof. We consider violation of (A2) in a linear model where association between X and $\boldsymbol{Z}$ is wholly due to unobserved confounding:

$$\begin{aligned} \boldsymbol{Z} &:= \boldsymbol{\epsilon_z}, \\ X &:= \epsilon_x, \\ Y &:= \theta^* X + \epsilon_y. \end{aligned}$$

We consider $d_X = 1$ and $d_{\boldsymbol{Z}} = 2$, take the exogenous variables to have a joint distribution $(\boldsymbol{\epsilon_z}, \epsilon_x, \epsilon_y) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_{\epsilon\epsilon}})$, where the covariance is given by the terms:

$$\boldsymbol{\Sigma_{\epsilon\epsilon}} := \begin{pmatrix} \eta_{Z_1}^2 & \rho_{Z_1 Z_2} \eta_{Z_1} \eta_{Z_2} & \rho_{Z_1 \epsilon_x} \eta_{Z_1} \eta_{\epsilon_x} & \rho_{Z_1 \epsilon_y} \eta_{Z_1} \eta_{\epsilon_y} \\ & \eta_{Z_2}^2 & \rho_{Z_2 \epsilon_x} \eta_{Z_2} \eta_{\epsilon_x} & \rho_{Z_2 \epsilon_y} \eta_{Z_2} \eta_{\epsilon_y} \\ & & \eta_{\epsilon_x}^2 & \rho_{\epsilon_x \epsilon_y} \eta_{\epsilon_x} \eta_{\epsilon_y} \\ & & & \eta_{\epsilon_y}^2 \end{pmatrix}.$$

We fix the following throughout the study:

$$\begin{aligned} \theta^* &:= 1, \\ \boldsymbol{\beta_x^*} &:= \mathrm{Cov}(X, \boldsymbol{Z}) = (2, -4), \\ \boldsymbol{\gamma_g^*} &:= \mathrm{Cov}(\epsilon_y, \boldsymbol{Z}) = (-2, 0.4), \end{aligned}$$

which, in turn, imply $\boldsymbol{\beta_y^*} = (0, 4.4)$.

D.1.1 Sweep Through Increasingly Uncertain Background Constraints

We consider three kinds of background constraints: (a) budget constraints, (b) an L_2-norm relaxation (a similar setting to Watson et al. (2024)), and (c) an L_1-norm relaxation (which is motivated by the relaxation in Kang et al. (2016), under the setting where point identification is not guaranteed):

(a) $\boldsymbol{\gamma_g} \in \boldsymbol{\Gamma}(\boldsymbol{\tau} = (\tau, 0.6), \boldsymbol{b} = (1, 2))$,
(b) $\|\boldsymbol{\gamma_g}\|_2 \leq \tau$,
(c) $\|\boldsymbol{\gamma_g}\|_1 \leq \tau$.

We adjust τ from 0 to 10 linearly across a grid of 101 simulation studies.

D.1.2 Randomized Parameters for the Covariance Matrix

We select random values for $\boldsymbol{\Sigma_{\epsilon\epsilon}}$ between simulations. The data generating process is constrained by two requirements: (i) the matrix $\boldsymbol{\Sigma_{\epsilon\epsilon}}$ must be positive definite, and (ii) the fixed values of $\boldsymbol{g^*}$ and $\boldsymbol{\beta_x^*}$, which demand:

$$(\gamma_g^*)_1 = \rho_{Z_1 \epsilon_y} \eta_{Z_1} \eta_{\epsilon_y}, \tag{9}$$

$$(\gamma_g^*)_2 = \rho_{Z_2 \epsilon_y} \eta_{Z_2} \eta_{\epsilon_y}, \tag{10}$$

$$(\beta_x^*)_1 = \rho_{Z_1 \epsilon_x} \eta_{Z_1} \eta_{\epsilon_x}, \tag{11}$$

$$(\beta_x^*)_2 = \rho_{Z_2 \epsilon_x} \eta_{Z_2} \eta_{\epsilon_x}. \tag{12}$$

We choose to generate $\boldsymbol{\Sigma_{\epsilon\epsilon}}$ by a rejection method. We sample the marginal variances according to:

$$\eta_{Z_1}, \eta_{Z_2}, \eta_{\epsilon_x}, \eta_{\epsilon_y} \overset{\text{i.i.d}}{\sim} \mathrm{Exp}(1),$$

and calculate the would-be ρ's from Eqs. (9) to (12). This enforces condition (ii). We then test whether the resultant $\boldsymbol{\Sigma_{\epsilon\epsilon}}$ is positive definite. If it is, (i) is satisfied and we perform the experiment; otherwise we resample

the η's from Exp(1). Fig. 3 shows the resulting feasible sets with plug-in $\hat{\beta}_\Phi$ and $\hat{\beta}_y$, calculated from a dataset of $N = 10000$ samples.

The feasible sets over θ in Fig. 3 were found using the efficient `budgetIV` algorithm for $d_\Phi = 1$, an implementation of which is provided in the supplement.

D.2 Nonlinear Simulation Study

In this experiment, we show another advantage of the budget background constraint approach over convex relaxations. Different values of the decision variables $\tilde{U}$ lead to different functions $\text{ATE}(\boldsymbol{x}; \boldsymbol{x_0})$.

D.2.1 Endogenous Equations

The candidate instruments are in the domain $\Omega_{\boldsymbol{Z}} = \{0, 1\}^{d_{\boldsymbol{Z}}}$, where $d_{\boldsymbol{Z}} = 6$; the exposure is a scalar in $\Omega_{\boldsymbol{X}} = [0, 1]$; and the outcome is a scalar in the full real line $\Omega_Y = \mathbb{R}$.

The ground truth structural equations are:

$$\boldsymbol{Z} := \boldsymbol{\epsilon_z}$$
$$X := f_x(\boldsymbol{Z}, \boldsymbol{\epsilon_x}),$$
$$Y := \boldsymbol{\theta}^* \cdot \boldsymbol{\Phi}^*(\boldsymbol{X}) + g_y(\boldsymbol{Z}, \boldsymbol{\epsilon_y}),$$

where the functions take the form:

$$f_x(\boldsymbol{Z}, \epsilon_x) = \text{Expit}(\epsilon_x - \boldsymbol{m} \cdot \boldsymbol{Z}),$$
$$g_y(\boldsymbol{Z}, \epsilon_y) = \lambda^{(A3)} \boldsymbol{Z} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{Z} + \lambda^{(A2)} \epsilon_y,$$
$$\Phi^*(X) = s_\Phi \left((X - 1/4)^2 - 1/16 \right),$$

and $\boldsymbol{\theta}^* = 1$.

The real vector $\boldsymbol{m} \in \mathbb{R}^{d_{\boldsymbol{Z}}}$ and the binary matrix $\boldsymbol{\Lambda} \in \{0, 1\}^{d_{\boldsymbol{Z}} \times d_{\boldsymbol{Z}}}$ are sampled randomly for each experiment. The components of $\boldsymbol{m}$ are i.i.d normal and the components of $\boldsymbol{\Lambda}$ are independent Bernoulli trials according to:

$$m_1, \ldots, m_{d_{\boldsymbol{Z}}} \overset{\text{i.i.d}}{\sim} \mathcal{N}(1, 4),$$

$$\Lambda_{ij} \sim \begin{cases} \mathcal{N}(0.9, 0.9) & i = j > 3, \\ \mathcal{N}(0.3, 0.3) & i \neq j \text{ and } i, j > 3, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that candidate instruments Z_1, Z_2, Z_3 all do not violate (A3).

The remaining parameters $s_\Phi, \lambda^{(A2)}, \lambda^{(A3)} \geq 0$ are restricted to be positive but are otherwise varied accross simulation settings (see Appx. D.2.3).

D.2.2 Generating the Exogenous Variables

We generate $(\boldsymbol{Z}, \epsilon_x, \epsilon_y)$ to have a block diagonal covariance structure where so that Z_1, Z_2 and Z_3 are valid instruments.

This is achieved by a Markov chain approach over binary Z with:

$$Z_1 \sim \text{Ber}(0.05),$$
$$Z_4 \sim \text{Ber}(0.05),$$
$$Z_{j+1} \mid Z_j = 1 \sim \text{Ber}(0.9),$$
$$Z_{j+1} \mid Z_j = 0 \sim \text{Ber}(0.05),$$

where $j \in \{2, 3, 5, 6\}$. Then the noise residuals are given by:

$$\epsilon_x = \boldsymbol{\gamma_{\epsilon_x}} \cdot \boldsymbol{Z} + \epsilon',$$
$$\epsilon_y = \boldsymbol{\gamma_{\epsilon_y}} \cdot \boldsymbol{Z} + \epsilon',$$

where:

$$
\begin{aligned}
(\gamma_{\epsilon_x})_1, \ldots, (\gamma_{\epsilon_x})_6 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \\
(\gamma_{\epsilon_y})_1, \ldots, (\gamma_{\epsilon_y})_3 &= 0, \\
(\gamma_{\epsilon_x})_4, \ldots, (\gamma_{\epsilon_x})_6 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \\
\epsilon' &\sim \mathcal{N}(0,1).
\end{aligned}
$$

D.2.3   Varying the Simulation Setting

The remaining parameters, $(s_\Phi, \lambda^{(\text{A2})}, \lambda^{(\text{A3})})$, are set to control the variance $\text{Var}(Y)$, signal to noise ratio SNR_Y, and the ratio of effects, defined by:

$$
R := \frac{\lambda^{(\text{A3})}}{\lambda^{(\text{A2})}} \frac{\|\text{Cov}(\boldsymbol{Z} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{Z}, \tilde{\boldsymbol{Z}})\|_2}{\|\text{Cov}(\epsilon_y, \tilde{\boldsymbol{Z}})\|_2},
$$

where $\tilde{\boldsymbol{Z}} := (Z_4, \ldots, Z_6)$.

Using the following shorthand:

$$
\begin{aligned}
P &:= \varphi(X) := \sqrt{1 + (X - 0.5)^2}, \\
\gamma_i &:= \boldsymbol{z}_i^\top \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{z}_i \\
r &:= \frac{\|\text{Cov}(\boldsymbol{\gamma}, \tilde{\boldsymbol{Z}})\|_2}{\|\text{Cov}(\epsilon_y, \tilde{\boldsymbol{Z}})\|_2}, \\
U &:= u(\boldsymbol{Z}, \epsilon_y) := \frac{r}{R} \gamma(\boldsymbol{Z}) + \epsilon_y,
\end{aligned}
$$

the remaining free parameters $s_\Phi, \lambda^{(\text{A2})}, \lambda^{(\text{A3})} \geq 0$ are expressed as:

$$
\begin{aligned}
s_\Phi^2 &= \frac{1}{\text{Var}\, P} \frac{\text{Var}\, Y}{1 + \frac{1}{\text{SNR}_Y}}, \\
\lambda^{(\text{A2})} &= \frac{s_\Phi \, \text{Cov}(P, U)}{\text{Var}\, U} \left(-1 + \frac{1}{2} \sqrt{1 + \frac{\text{Var}\, U}{s_\Phi^2 (\text{Cov}(P, U))^2} \frac{\text{Var}\, Y}{1 + \text{SNR}_Y}} \right), \\
\lambda^{(\text{A3})} &= \frac{R}{r} \lambda^{(\text{A2})}.
\end{aligned}
$$

We choose to fix $\text{Var}(Y) := 10$ (this is simply a choice of the scale with which we choose to measure Y). We produce a three-by-three grid of experiments with the values of $\text{SNR}_Y = 1/10, 1/5, 1$ and $R = 1/2, 1, 2$.

The remaining parameters are fixed to approximately fit these constraints by a post-hoc method using plug-in empirical estimates: $\widehat{\text{Var}}(P), \widehat{\text{Var}}(U), \widehat{\text{Cov}}(P, U)$ and a plug-in ratio estimate:

$$
\hat{r} := \frac{\|\widehat{\text{Cov}}(\boldsymbol{Z} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{Z}, \tilde{\boldsymbol{Z}})\|_2}{\|\widehat{\text{Cov}}(\epsilon_y, \tilde{\boldsymbol{Z}})\|_2},
$$

where a dataset of 5×10^5 samples is used to construct the estimates.

D.2.4   Results from Main Text

With the same $N = 5 \times 10^5$ sample dataset, we run an implementation of `budgetIV`, included in the supplement, with feasible search space $\boldsymbol{\Gamma}(\tau = 0, b = 3)$. Fig. 6 shows the results for the full simulation grid.
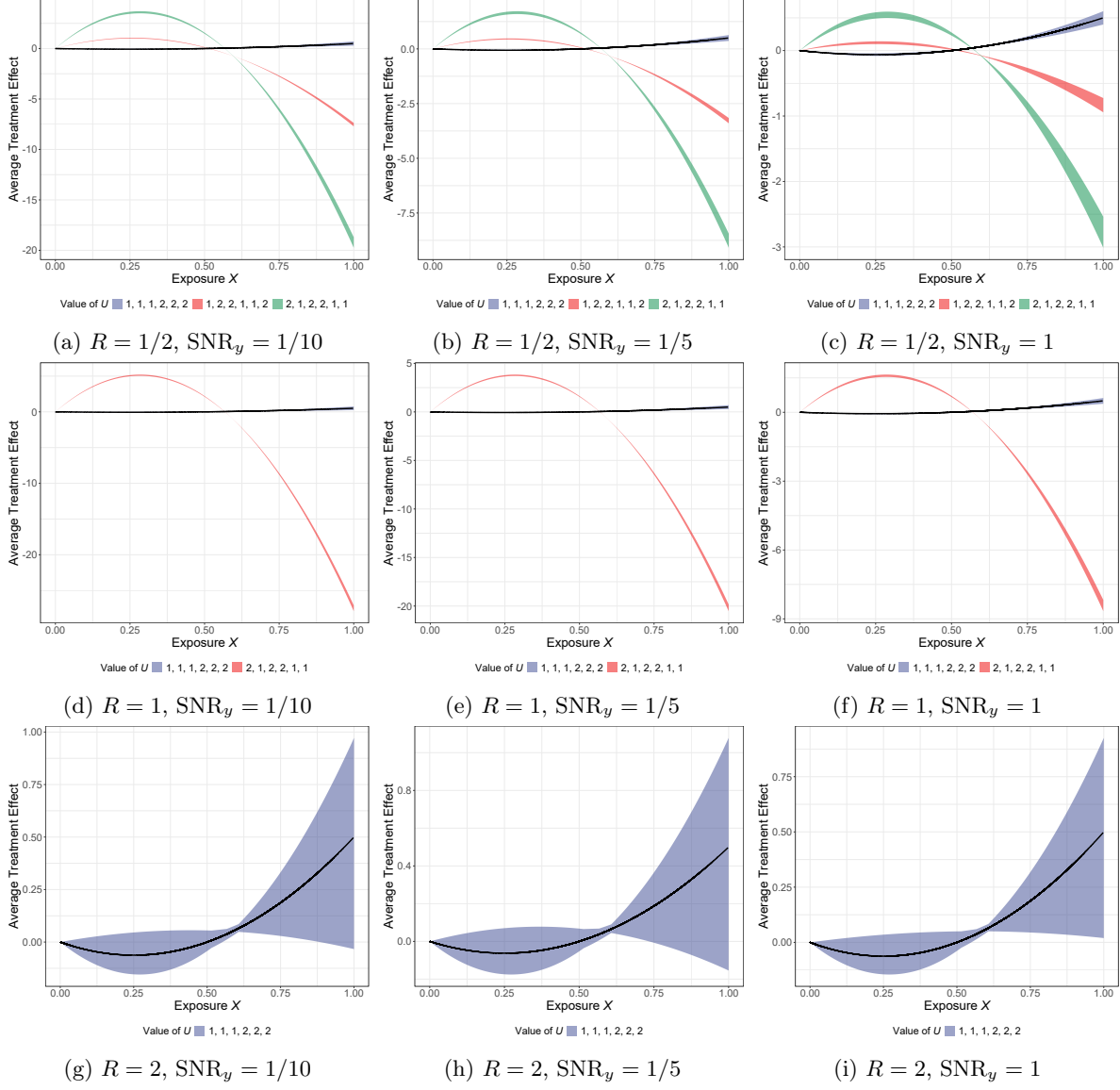
Figure 6: **Simulation grid for the experiment discussed in Appx. D.2**. Some settings (a, b) result in a feasible set in which the ATE is qualitatively very different for each plausible $\tilde{\boldsymbol{U}}$. In other settings (g–i) $\boldsymbol{U}^*$ is identified exactly. $\Sigma_{\boldsymbol{b}}^{(\max)} = \binom{6}{3} = 20$, so budgetIV significantly reduced the space of plausible $\tilde{\boldsymbol{U}}$ in this experiment.
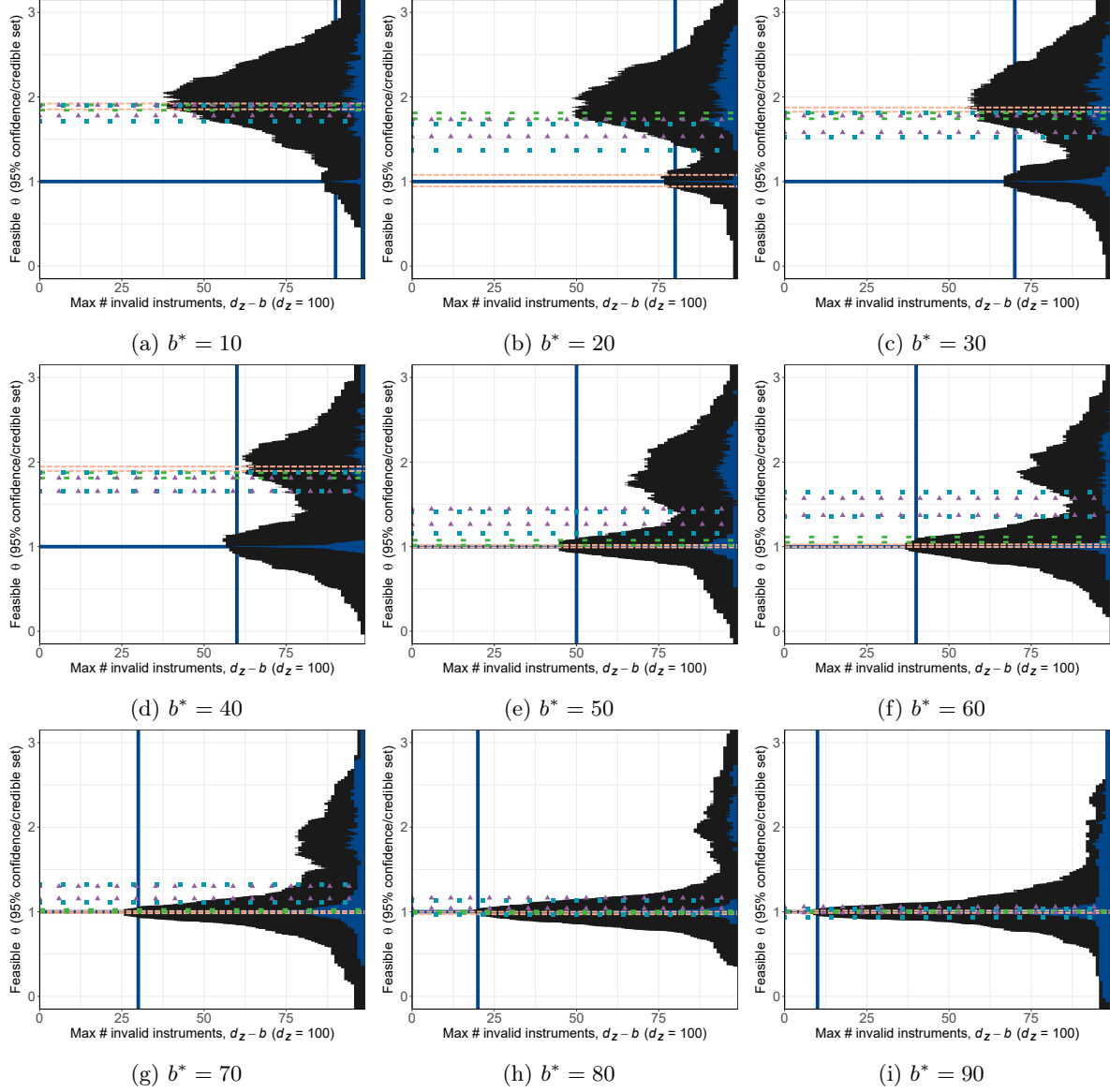
Figure 7: **Simulation grid for the experiment discussed in Appx. D.3**. Each subfigure corresponds to a different ground truth number of valid IVs b^*. Finite sample confidence sets using `budgetIV_scalar` and **Oracle** are shown for the constraints $\mathbf{\Gamma}(\tau = 0, b)$ where b is varied along the horizontal axis. Confidence intervals for the benchmarking methods **MR-Egger**, **MR-Median**, **MBE** and **IVW**, under which b is not an adjustable parameter, are labeled for each experiment. Sub-figures (a) through (e) show variability in the confidence intervals of **MBE** when valid IVs are a minority—despite the required modal assumption of holding for each ground truth model. The approach **MASSIVE**, included in Fig. 5 in the main text, takes significantly more computational resources than the the methods included in this simulation grid.
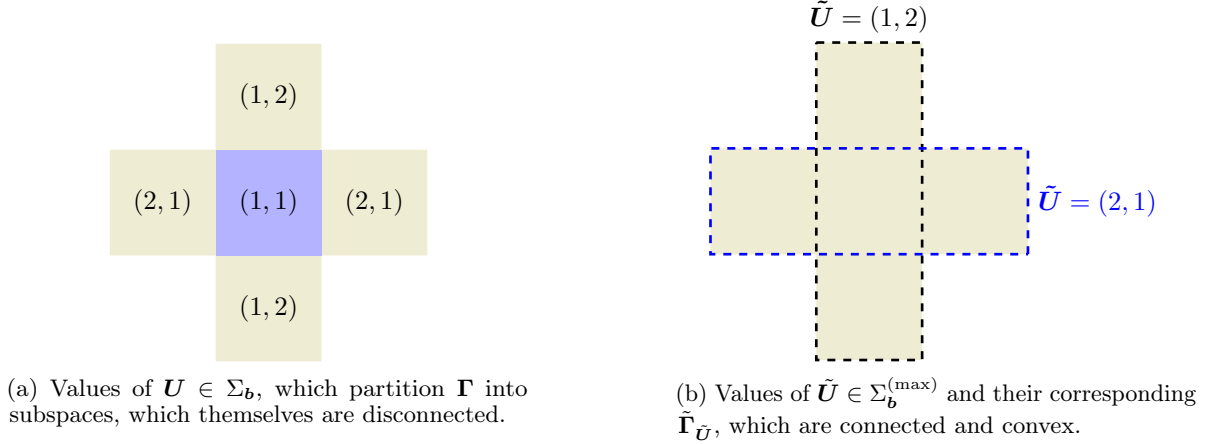
(a) Values of $\boldsymbol{U} \in \Sigma_{\boldsymbol{b}}$, which partition $\boldsymbol{\Gamma}$ into subspaces, which themselves are disconnected.

(b) Values of $\tilde{U} \in \Sigma_{\boldsymbol{b}}^{(\max)}$ and their corresponding $\tilde{\boldsymbol{\Gamma}}_{\tilde{U}}$, which are connected and convex.

Figure 8: A comparison between representing the search space $\boldsymbol{\Gamma}(\boldsymbol{\tau} = (\tau_1, \tau_2), \boldsymbol{b} = (1, 2))$ as a disjoint union over $\boldsymbol{\Gamma}_{\boldsymbol{U}}$ or a union over $\tilde{\boldsymbol{\Gamma}}_{\tilde{U}}$.

D.3 Linear Many Candidate IV Study

Simulation setting We performed a grid of simulations using the following linear SEM with (A3) violation among a varying proportion of the 100 candidate IVs:

$$
\begin{aligned}
\boldsymbol{Z} &:= \boldsymbol{\epsilon_z}, \\
X &:= \boldsymbol{\delta} \cdot \boldsymbol{Z} + \epsilon_x. \\
Y &:= \theta^* X + \boldsymbol{\gamma} \cdot \boldsymbol{Z} + \epsilon_y, \\
b^* &:= \|\boldsymbol{\gamma}_g^*\|_0 \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}, \\
\theta^* &:= 1.
\end{aligned}
$$

We set $(\boldsymbol{\epsilon_z}, \epsilon_x, \epsilon_y)$ to a multivariate normal centered at $\boldsymbol{0}$ with no correlation except $\rho := \mathrm{Corr}(\epsilon_x, \epsilon_y)$ drawn uniformly from $[-1, 1]$. The marginal standard deviations are drawn i.i.d. from $\mathrm{Exp}(1)$. The first b^*-many entries of $\boldsymbol{\gamma}$ are set to 0 and the remaining entries, as well as the entries of $\boldsymbol{\delta}$, are samples i.i.d. from the uniform distribution $\mathcal{U}[1, 2]$.

A two-sample approach was applied with $N_x = 1 \times 10^6$ samples used to sample data to generate the summary statistics $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}} := \widehat{\mathrm{Cov}}(\boldsymbol{Z}, X)$ and $N_y = 1 \times 10^5$ samples used to generate $\hat{\boldsymbol{\beta}}_{\boldsymbol{y}} := \widehat{\mathrm{Cov}}(\boldsymbol{Z}, X)$. This was to model the typical occurrence that $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}}/\mathrm{SE}(\hat{\boldsymbol{\beta}}_{\boldsymbol{x}}) \gg \hat{\boldsymbol{\beta}}_{\boldsymbol{y}}/\mathrm{SE}(\hat{\boldsymbol{\beta}}_{\boldsymbol{y}})$ (SE is the empirical standard error) because candidate instruments are selected if they have a strong association with X (see Bowden et al. (2016b)).

This simulation setting reflects common modeling assumptions that are applied in the original experiments for each of the benchmark methods.

Results Results for the case $b^* = 30$ were interpreted in Sect. 5, including the caption to Fig. 5. Further interpretation, including to the stability of the MBE estimator is given in the caption to Fig. 7.

E SEARCHING THROUGH THE BUDGET CONSTRAINTS

The number of unique decision variables $|\Sigma_{\boldsymbol{b}}|$ is given by the number of assignments $\boldsymbol{U} \in [K+1]^{d_{\boldsymbol{Z}}}$ for which at least b_1 components of $\boldsymbol{U}$ are equal to 1, at least b_2 components are no greater than 2 and so on until exactly $b_{K+1} := d_{\boldsymbol{Z}}$ components.

Suppose exactly d_1 components are equal to 1, exactly $d_2 - d_1$ are equal to 2 and so on. This number of combinations with this property is equal to the multinomial coefficient $\binom{d_{\boldsymbol{Z}}}{d_1, (d_2 - d_1), \dots, (d_{K+1} - d_K)}$.

Given that we require $d_i \geq b_i$ for all $i \in [K+1]$, we can write the following:

$$|\Sigma_{\boldsymbol{b}}| = \sum_{d_1=b_1}^{d_2} \cdots \sum_{d_K=b_K}^{d_{\boldsymbol{Z}}} \frac{d_{\boldsymbol{Z}}!}{\prod_{\ell=1}^{K+1}(d_\ell - d_{\ell-1})!}.$$

By comparison, $|\Sigma_{\boldsymbol{b}}^{(\text{max})}|$ count $\tilde{\boldsymbol{U}} \in [K+1]^{d_{\boldsymbol{Z}}}$ for which exactly b_1 components equal 1, $b_2 - b_1$ components equal 2 and so on. Therefore, denoting $b_0 := 0$, we have:

$$|\Sigma_{\boldsymbol{b}}^{(\text{max})}| = \frac{d_{\boldsymbol{Z}}!}{\prod_{\ell=1}^{K+1}(b_\ell - b_{\ell-1})!}.$$

The ratio gap between these sizes is given exactly by:

$$\frac{|\Sigma_{\boldsymbol{b}}|}{|\Sigma_{\boldsymbol{b}}^{(\text{max})}|} = \sum_{d_1=b_1}^{d_2} \cdots \sum_{d_K=b_K}^{d_{\boldsymbol{Z}}} \frac{\prod_{\ell=1}^{K+1}(b_\ell - b_{\ell-1})!}{\prod_{\ell=1}^{K+1}(d_\ell - d_{\ell-1})!} > 1.$$

Depending on the scaling of $d_{\boldsymbol{Z}}$ and $\boldsymbol{b}$, asymptotic gap can be $\mathcal{O}(1)$ or as extreme as $\mathcal{O}(d_{\boldsymbol{Z}}^K)$.

Fig. 8 shows that the subsets $\boldsymbol{\Gamma_U} \subset \boldsymbol{\Gamma}(\boldsymbol{\tau}, \boldsymbol{b})$ for $\boldsymbol{U} \in \Sigma_{\boldsymbol{b}}$ are disconnected, while $\tilde{\boldsymbol{\Gamma}}_{\tilde{\boldsymbol{U}}} \subset \boldsymbol{\Gamma}$ for $\tilde{\boldsymbol{U}} \in \Sigma_{\boldsymbol{b}}^{(\text{max})}$ are cuboids in general. Therefore, it is clear that testing for intersection between $h(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\Gamma}}_{\tilde{\boldsymbol{U}}}$ is more efficient than searching through the decision variables $\boldsymbol{U}$ directly.

F SUMMARY OF RELAXED IV METHODS

Paper	$d_X > 1$	$d_Z > 1$	Cont.	Nonlin.	2-sample	Inference	Violation	Feasible γ_g
Conley et al. (2012)	✓	✓	✓	✗	✗	✓	(A3)	Convex set
Nevo and Rosen (2012)	✓	✓	✓	✗	✗	✓	(A2)	Convex set
Ramsahai (2012)	✗	✗	✗	✓	✓	✗	(A2) ∨ (A3)	N/A
Bowden et al. (2015)	✗	✓	✓	✗	✓	✓	(A3)†	$d_Z \to \infty$
Kolesár et al. (2015)	✗	✓	✓	✗	✗	✓	(A3)†	$d_Z \to \infty$
Bowden et al. (2016a)	✗	✓	✓	✗	✗	✓	(A3)	Sparse
Kang et al. (2016)	✗	✓	✓	✗	✗	✗	(A3)*	Sparse
Silva and Evans (2016)	✗	✓	✗	✓	✓	✓	(A2) & (A3)	N/A
Hartwig et al. (2017)	✗	✓	✓	✗	✓	✓	(A3)	Mode zero
Guo et al. (2018)	✗	✗	✓	✓	✗	✓	(A3)	Mode zero
Windmeijer et al. (2019)	✗	✓	✓	✗	✗	✓	(A3)*	Sparse
Shapland et al. (2019)	✗	✓	✓	✗	✗	✓	(A3)*	L_0-norm
Bucur et al. (2020)	✗	✓	✓	✗	✓	-	(A3)*	L_0-norm
Kang et al. (2020)	✗	✓	✓	✗	✗	✓	(A2) & (A3)	Point
Kuang et al. (2020)	✗	✓	✗	✗	✗	✗	(A2)* & (A3)*	Point
Hartford et al. (2021)	✗	✓	✓	✓	✗	✗	(A3)*	Mode zero
Vancak and Sjölander (2023)	✗	✗	✓	✓	✗	✓	(A2) & (A3)	Point
Xue et al. (2023)	✗	✓	✓	✗	✓	-	(A2) & (A3)	L_0-norm
Watson et al. (2024)	✗	✓	✓	✗	✓	✓	(A3)	Convex set
`budgetIV`	✓	✓	✓	✓	✓	✓	(A2) & (A3)	Star domain

Table 1: **Summary of the constraints and affordances of some notable relaxed IV methods.** We exclude approaches that adhere to classical IV assumptions but provide nonparametric partial identification bounds as out of scope (e.g., Balke and Pearl (1997); Kilbertus et al. (2020); Levis et al. (2023)). We do the same with more generic algorithms not specifically designed for the IV setting (e.g., Hu et al. (2021); Duarte et al. (2023); Padh et al. (2023)). The columns denote, in order, if the method allows for: (1) multidimensional exposures; (2) multiple candidate instruments; (3) continuous data; (4) nonlinear structural equations; (5) summary statistic input; and (6) statistical inference (where **-**corresponds to point estimators in partially identifiable settings). The final two columns indicate (7) which IV assumptions (if any) may be violated and (8) the assumed geometry of the feasible region. The * in column (7) indicates that only *some* candidate IVs are allowed to violate starred assumptions (typically fewer than 50%), while † indicates that the mechanism behind a candidate IV's (A3) violation must be independent to the effect of the candidate on the exposure. The N/A entries in column (8) correspond to fully nonparametric models in which γ_g is undefined.