Stochastic Causal Programming for Bounding Treatment Effects

https://arxiv.org/pdf/2202.10806.pdf

joint work with
Jakob Zeitler, David Watson, Matt Kusner,
Ricardo Silva, Niki Kilbertus

Kirtan Padh

HELMHOLTZAI

TUM

UCL

KING'S College LONDON

Motivation

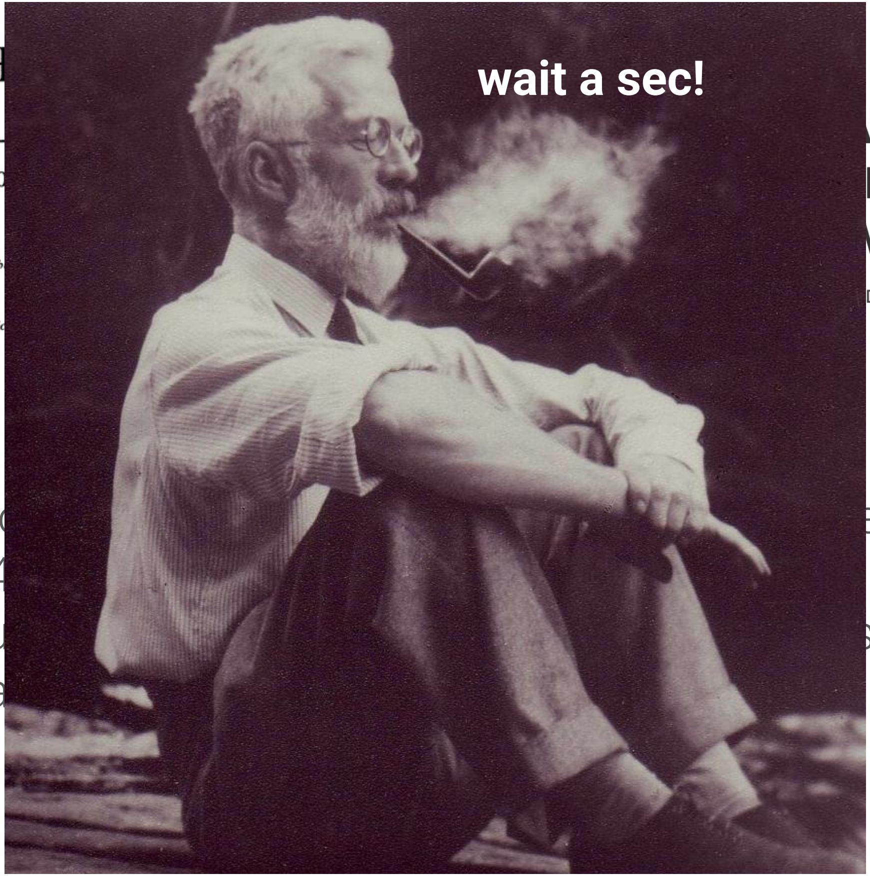
There was “a lot of correlation”

BRITISH

SMO

Memb

Professor of Medical Statistic



ATIONSHIP BETWEEN HUMAN SMOKING
ND DEATH RATES
W-UP STUDY OF 187,766 MEN

D.; Daniel Horn, Ph.D.

- 3 c ers, 56 were heavy smokers
- 14 . 23.9% other cancer patients
- su st (all 36 who died of lung ca
- m

Unobserved confounding

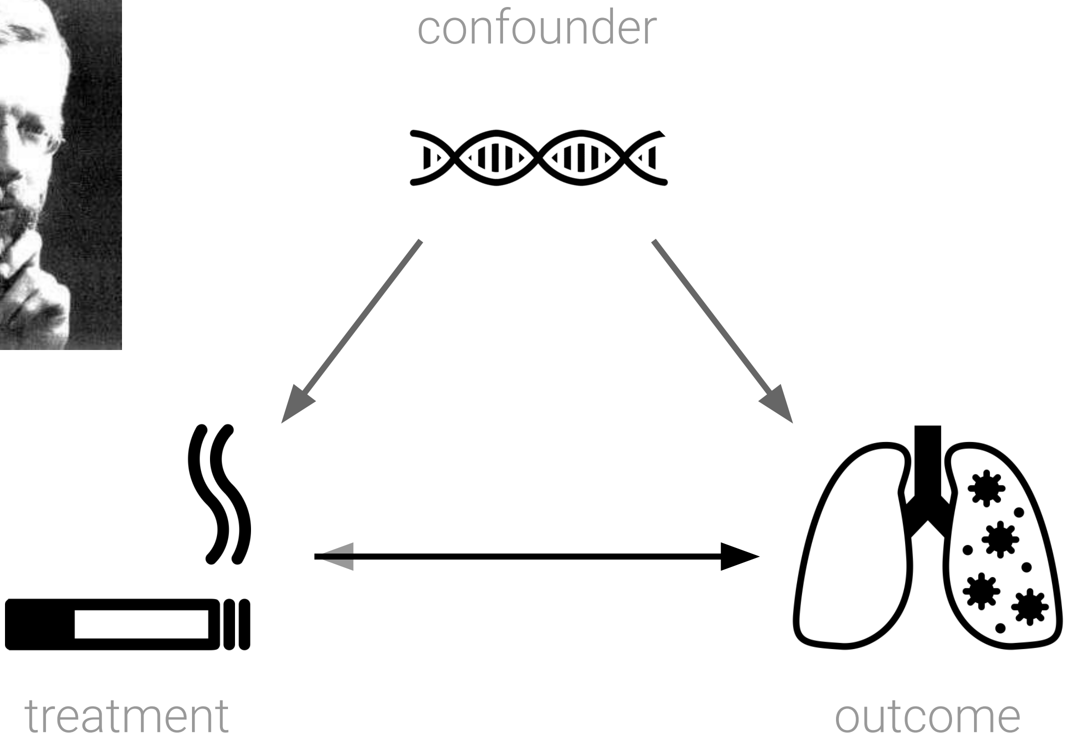


Image credit (from Noun Project):
priyanka, Andrew Nielsen, mungang kim

Naive ML approach fails

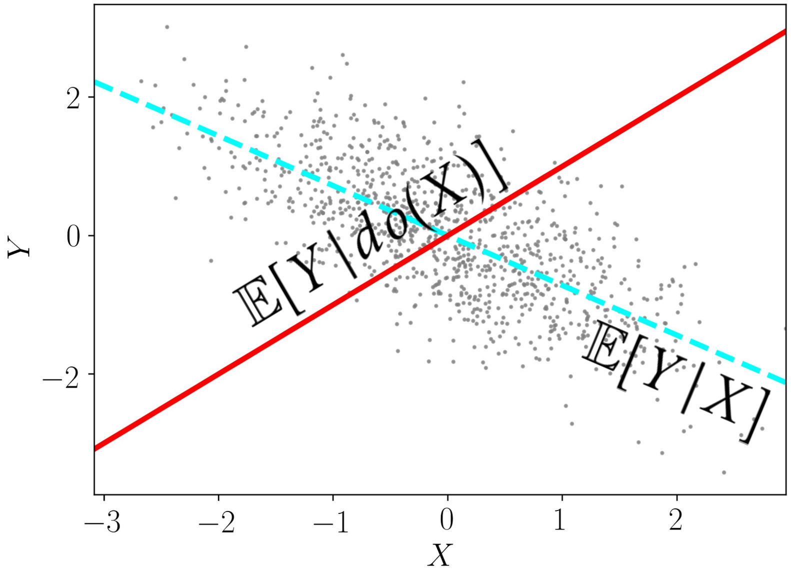
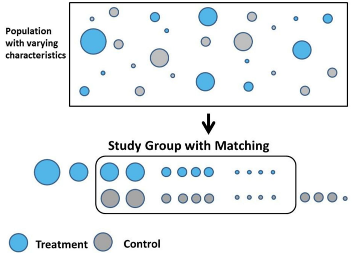


Image credit:
Niki Kilbertus

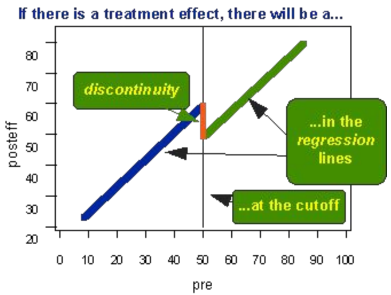
How do we estimate causal effects from observational data?

Some of the ways

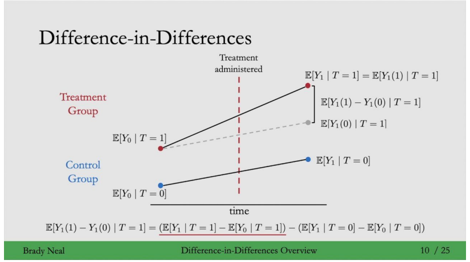


Covariate matching [1]

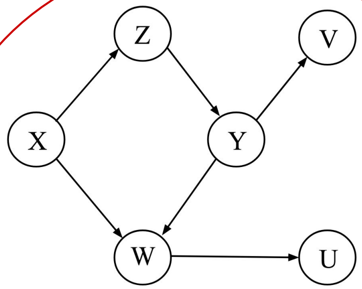
We are here!



Regression discontinuity [3]



Difference-in-differences [2]



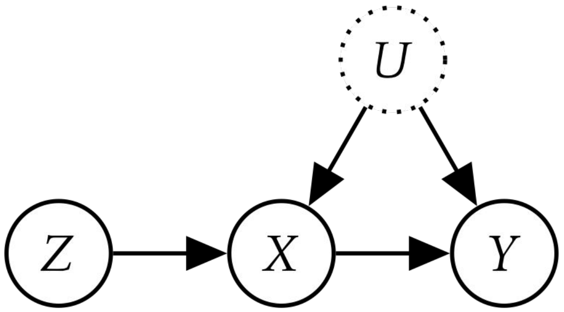
Structural assumptions [4]

Image credits:
1 https://statistikakademin.se, 2 Brady Neal, 3 William M.K. Trochim, 4 Fabian Dablander

Structural assumptions

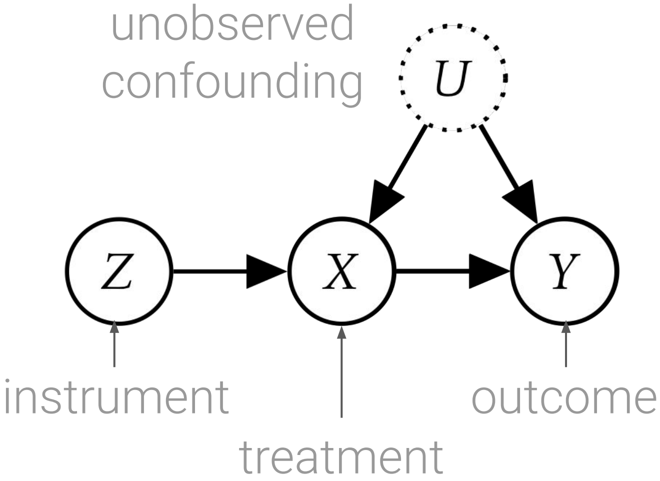
A known causal graph (DAG) with hidden confounding.

The method applies to general graphs



We focus on the **Instrument Variable (IV) setting**

IV: Identifiability (through additive noise)



(a) Z influences X $\quad Z \not\perp\!\!\!\perp X$

(b) Z is independent of U $\quad Z \perp\!\!\!\perp U$

(c) Z only influences Y via X $\quad Z \perp\!\!\!\perp Y \mid \{X, U\}$

assume: $Y = f(X) + e_Y \quad \textbf{with} \quad \mathbb{E}[e_Y] = 0$

$$\boxed{\mathbb{E}[Y \mid z]} = \mathbb{E}[f(X) + e_Y \mid z] = \mathbb{E}[f(X) \mid z] = \int \boxed{f(x)}\,\boxed{p(x \mid z)}\,dx$$

identifiable

unique under
mild conditions

identifiable

Even in the IV setting, conditions for identifiability are still (too?) strong

But … bility

Assur … IV setting

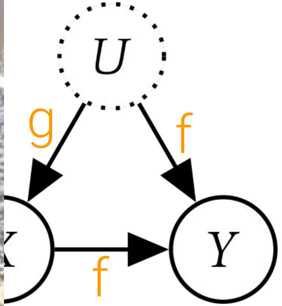
Point ic	entifiability

Addit …
$Y = f(X) + e_Y$ …

… nite number of
[Gunsilius, 2019]

Mon …
$\Pr(x \mid z) \leq \Pr$ …

and …

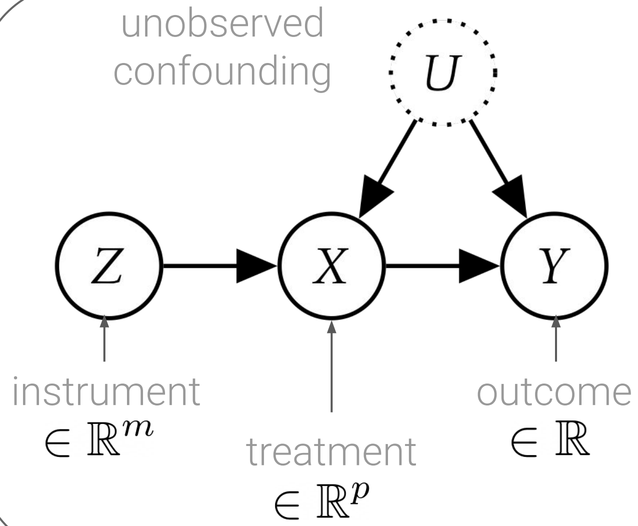


The story so far

- A zoo of point identifia 1930s.
- But less work on partia
  - Binary variables, I
  - Finite variables, g
  - Discrete variables 21]
  - Scalar variables [ ]
  - High-dimensional]

Problem formulation

General problem formulation



Assumptions

(a) Z influences X $\quad Z \not\perp\!\!\!\perp X$

(b) Z is independent of U $\quad Z \perp\!\!\!\perp U$

(c) Z only influences Y via X $\quad Z \perp\!\!\!\perp Y \mid \{X, U\}$

$$X = g(Z, U) \qquad Y = f(X, U)$$

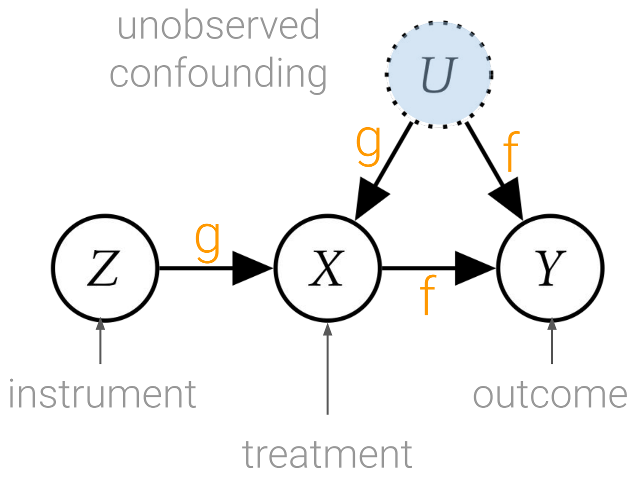
non-linear, non-additive

Goal - partial identification

For any x^* compute lower and upper bounds on the causal effect

$$\mathbb{E}[Y \mid do(x^\star)]$$

General problem formulation as optimization



optimize over “all” distributions

$$X = g(Z, U) \qquad Y = f(X, U)$$

optimize over “all” functions

Goal

among all possible {*g*, *f*} and distributions over *U*
that reproduce the observed densities {p(y | x, z), p(x | z)},
estimate the min and max expected outcomes under intervention

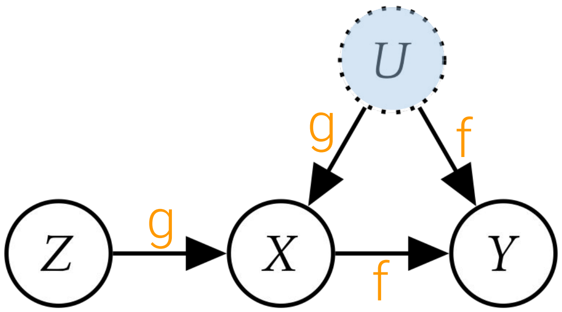
A causal mathematical program

Lower bound < Causal Effect < Upper bound

$$
\begin{aligned}
\underset{\eta_\theta, w_\psi}{\text{min / max}} \quad & o(S) \\
\text{subject to} \quad & \text{dist}(p_S, \hat{p}) \leq \epsilon \,, \\
& \text{structural constraints,}
\end{aligned}
$$

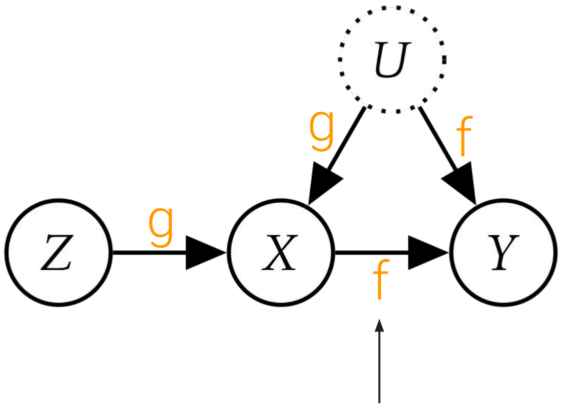
Cannot have no restrictions on f and g

- Without any restrictions on functions and distributions:
  effect is not identifiable and average treatment effect bounds are vacuous
  [Pearl, 1995; Bonet, 2001; Gunsilius 2018]
- Mild assumptions suffice for meaningful bounds:
  f and *g* have a finite number of discontinuities [Gunsilius, 2019]
- Rest of the talk: **operationalize the optimization**



Our practical approach

Response functions I [Balke & Pearl, 1994]



ultimately, we care about this functional relation

- Each value of U fixes a functional relation $X \rightarrow Y$
- Collect the set of all resulting functions $\{f_u\}$
- Identify values of u that result in the same f_u and assign a unique index r

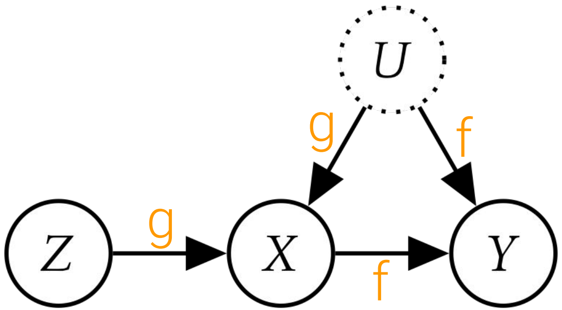
$$Y = f(X, U) = \lambda_1 X + \lambda_2 X U_1 + U_2$$

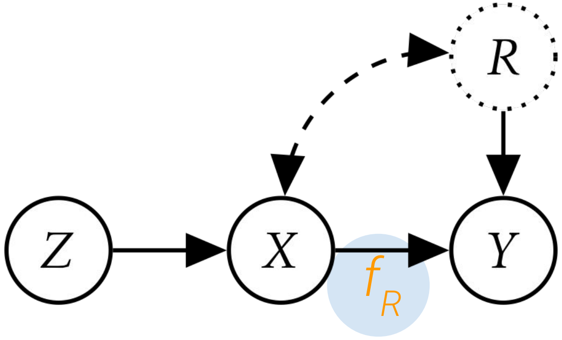
$$f(x, u) = \lambda_1 x + \lambda_2 x \quad \text{for} \quad u_1 = 1, u_2 = 0$$

$$f_r(x) = (\lambda_1 + \lambda_2) x \quad \text{where} \quad r \text{ is an alias for } (1, 0)$$

→ Instead of a potentially multivariate distribution over confounders U directly, we can think of a distribution R over functions f: $X \rightarrow Y$

Response functions II





choose convenient function spaces

find convenient representation of U from which we can sample

find convenient representation of distributions over response functions

Parameterizing response functions

We choose a simple parameterization

$$f_r(x) := f_{\theta_r}(x) \quad \text{for} \quad \theta \in \Theta \subset \mathbb{R}^K$$

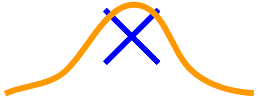
For simplicity, work with linear combination of (non-linear) basis functions:

$$f_\theta(x) = \sum_{k=1}^{K} \theta_k \psi_k(x) \text{ for basis functions } \{\psi_k : \mathbb{R}^p \to \mathbb{R}\}_{k \in [K]}$$

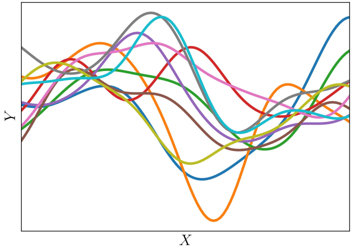
polynomials

neural networks

...

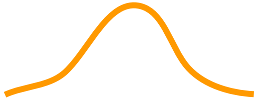


θ



$f_\theta : X \to Y$

Parameterizing the distribution over θ



$p_{\mathcal{M}}(\theta)$ implies a causal model, and a distribution $\hat{p}_{\mathcal{M}}(x, y, z)$

Goal

Optimize over distributions $p_{\mathcal{M}}(\theta)$ such that

$$\hat{p}_{\mathcal{M}}(x, y, z) \text{ is close to the observed distribution } p(x, y, z)$$

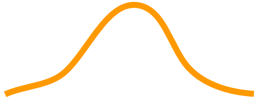
Ideally

low variance Monte-Carlo gradient estimation

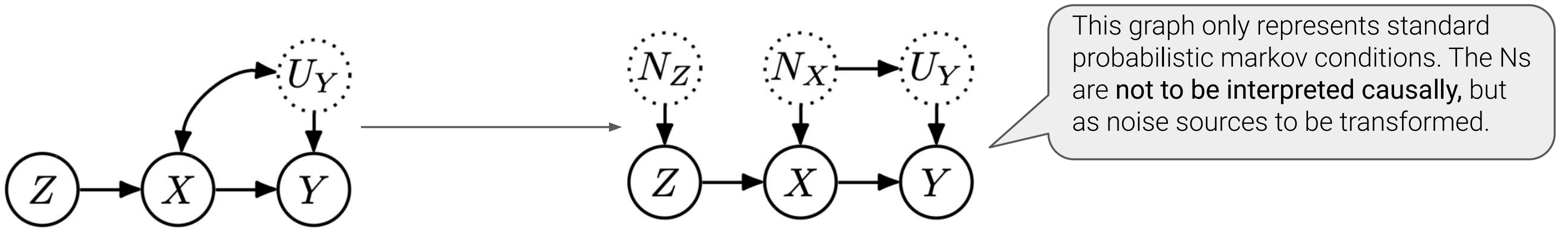
differentiable sampling

again, assume parametric form of $p_{\mathcal{M}}(\theta)$

$$p_\eta(\theta) \quad \text{with} \quad \eta \in \mathbb{R}^d$$



The parametrization in practice



$$\boldsymbol{\theta} \mid N_X \sim p_\eta\left(\cdot\, ;\; \mu_{\eta_0}\left(N_X\right), \Sigma_{\eta_1}\left(N_X\right)\right)$$

The distribution is defined up to mean and covariance functions.

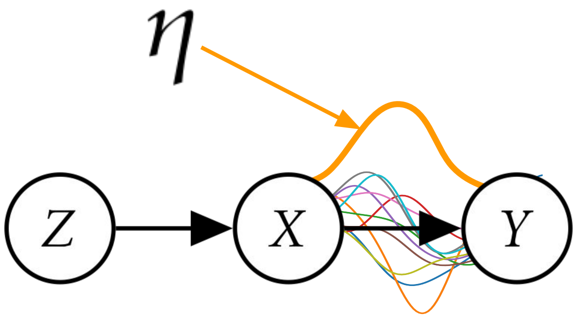
Optimization Parameters: $\eta = (\eta_0, \eta_1)$
μ_{η_0} and Σ_{η_1} are small neural nets

A causal mathematical program

$$\min_{\eta_\theta, w_\psi} / \max \quad o(S)$$
$$\text{subject to} \quad \text{dist}(p_S, \hat{p}) \leq \epsilon\,,$$
$$\text{structural constraints,}$$

Objective: ATE [obj] Optimising causal effect, e.g. ATE	**Constraint: Data [c-data]** Matching the observed data distribution	**Constraint: Structure [c-struct]** Graphical assumptions
Estimating $\mathbb{E}[Y \mid do(x^*)]$	Matching p(x | z) and p(y | x, z)	Baked into the model

Objective function



objective

$$\min_\eta / \max_\eta \mathbb{E}[Y \,|\, do(x^\star)] = \min_\eta / \max_\eta \int f_\theta(x^\star) p_\eta(\theta)\, d\theta$$

$$\boxed{= \psi_Y(x^\star)^\top \mathbb{E}_{N_X}[\mu_{\eta_0}(N_X)]}$$

A causal mathematical program

$$\min_{\eta_\theta, w_\psi} / \max \quad o(S)$$

$$\text{subject to} \quad \text{dist}(p_S, \hat{p}) \leq \epsilon \,,$$

$$\text{structural constraints,}$$

Objective: ATE [obj] Optimising causal effect, e.g. ATE	**Constraint: Data [c-data]** Matching the observed data distribution	**Constraint: Structure [c-struct]** Graphical assumptions
Estimating $\mathbb{E}[Y \vert do(x^*)]$	Matching p(x | z) and p(y | x, z)	Baked into the model

Match p(x | z)

Identified from data and manually fixed once up front.
Implemented as an invertible **conditional normalizing flow**.

$$x = h_z(n)$$

Note: Given x_i, z_i, we can uniquely determine $n_i = h_{z_i}^{-1}(x_i)$

Match p(y | x, z)

Match the first two moments at a representative, finite set of points from p(x, z)

For the IV, the constraints are then in closed form.

$$\mathbb{E}[f_\theta(x_j) \mid x_j, z_j] = \psi(x_j)^\top \mu_{\eta_0}(n_j)$$

$$\mathbb{E}[f_\theta^2(x_j) \mid x_j, z_j] =$$
$$\psi(x_j)^\top \mu_{\eta_0}(n_j) \left(\Sigma_{\eta_1}(n_j) + \mu_{\eta_0}(n_j)\mu_{\eta_0}^\top(n_j)\right) \psi(x_j)$$

$$j \in \{1, 2, \ldots, D\}$$

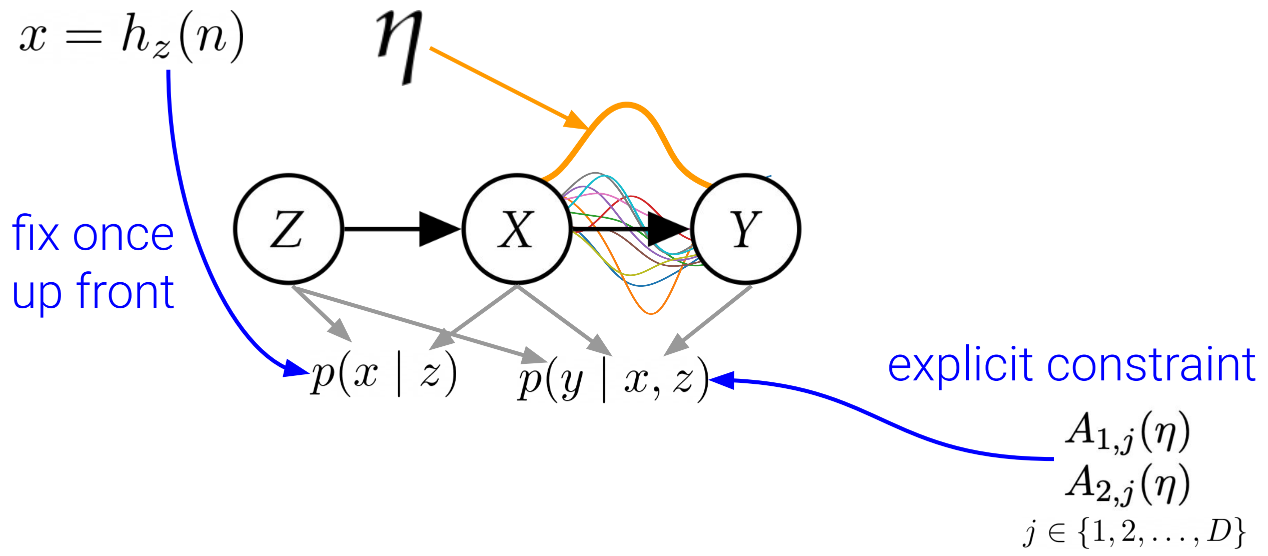
A random subsample from the data
(The ‘stochastic’ in stochastic causal programming)

A causal mathematical program

$$\begin{aligned} \min_{\eta_\theta, w_\psi} / \max \quad & o(S) \\ \text{subject to} \quad & \text{dist}(p_S, \hat{p}) \leq \epsilon \,, \\ & \text{structural constraints,} \end{aligned}$$

Objective: ATE [obj] Optimising causal effect, e.g. ATE	**Constraint: Data [c-data]** Matching the observed data distribution	**Constraint: Structure [c-struct]** Graphical assumptions
Estimating $\mathbb{E}[Y \mid do(x^*)]$	Matching p(x | z) and p(y | x, z)	Baked into the model

Intermediate overview



objective

$$\min_{\eta} / \max_{\eta} \mathbb{E}[Y \mid do(x^\star)]$$

The final optimization problem

Lower bound < Causal Effect < Upper bound

Precompute once up front from data

The final values

$$\min_{\eta_\theta, w_\psi} / \max \quad o(S)$$
$$\text{subject to} \quad \text{dist}(p_S, \hat{p}) \leq \epsilon\,,$$
$$\text{structural constraints},$$

Constraints enforced using the **augmented Lagrangian** (Nocedal and Wright, 2006)

Objective: [obj] ATE	**Constraint: Data [c-data]** Matching the observed data distribution	**Constraint: Structure [c-struct]** Graphical assumptions
$\psi_Y(x^\star)^\top \mathbb{E}_{N_X}[\mu_{\eta_0}(N_X)]$	$x = h_z(n)$ (precompute $\{n_j\}_{j\in[D]}$) $\text{LHS}_{j,l} = \mathbb{E}[\phi_l(Y) \mid x_j, z_j]$ $\text{RHS}_{j,l}(\eta) = A_{j,l}$ $\text{LHS}_{j,l} = \text{RHS}_{j,l}(\eta)$ $\forall l \in [2]$ and $j \in [D]$	Baked into the model

Empirical results

Choices of response functions

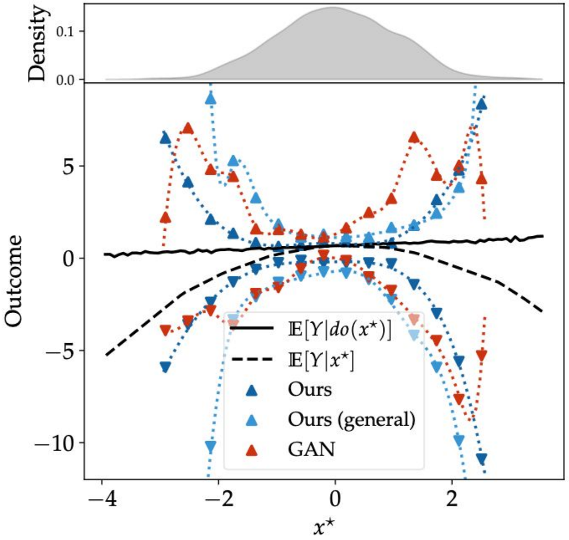
$$f_\theta(x) = \sum_{k=1}^{K} \theta_k \psi_k(x) \text{ for basis functions } \{\psi_k : \mathbb{R}^p \to \mathbb{R}\}_{k\in[K]}$$

Polynomials

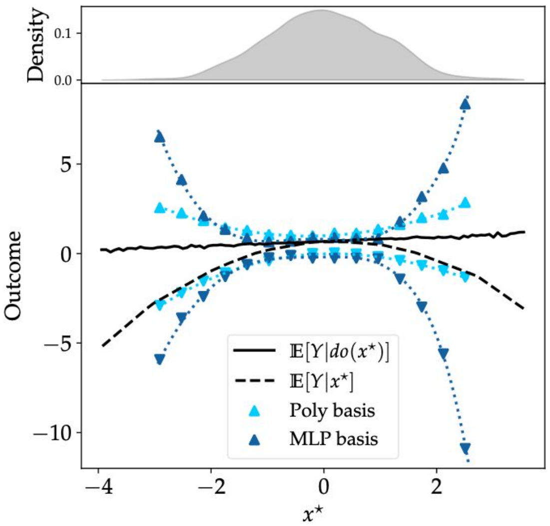
Neural network
Train a small fully connected network on observed data $X \rightarrow Y$ and take activations of last hidden layer as basis functions.

For visualization: All interventions are along a single axis for multi-dimensional treatments

Linear 2-D treatment, strong confounding

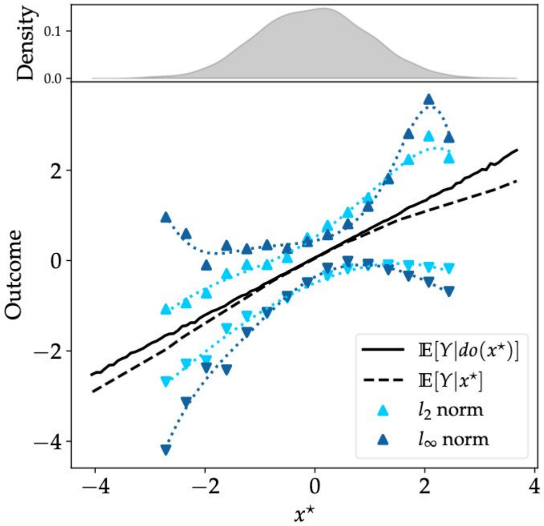


Ours seems more stable



Stronger assumptions, stronger inference

Linear 2-D treatment, weak confounding



$$\sum_{j,l} (\mathsf{LHS}_{j,l} - \mathsf{RHS}_{j,l})^2 = 0$$

Single constraint [c-data]

Hu, Y., Wu, Y., Zhang, L., & Wu, X. (2021, May). A generative adversarial framework for bounding confounded causal effects. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 13, pp. 12104-12112).

Takeaways

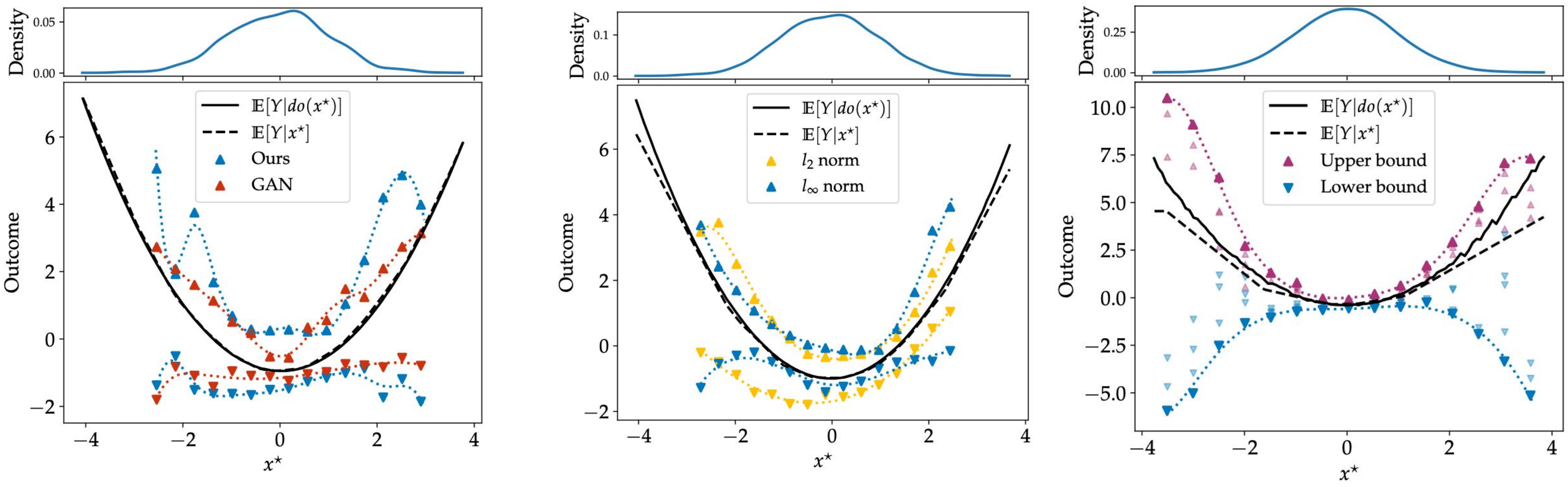
- Partial identification is hard for continuous high-dimensional variables.
- We were able to craft a framework that is
    - **flexible**
    - **needs minimal Monte-Carlo estimations** in the IV and leaky mediator settings
    - allows the user the choice of a **spectrum of assumption strength**

See the paper for extensions to more general settings.

Quadratic 3-D treatment, weak confounding

Quadratic 2-D treatment, weak confounding

Quadratic scalar treatment, weak confounding



Thank you