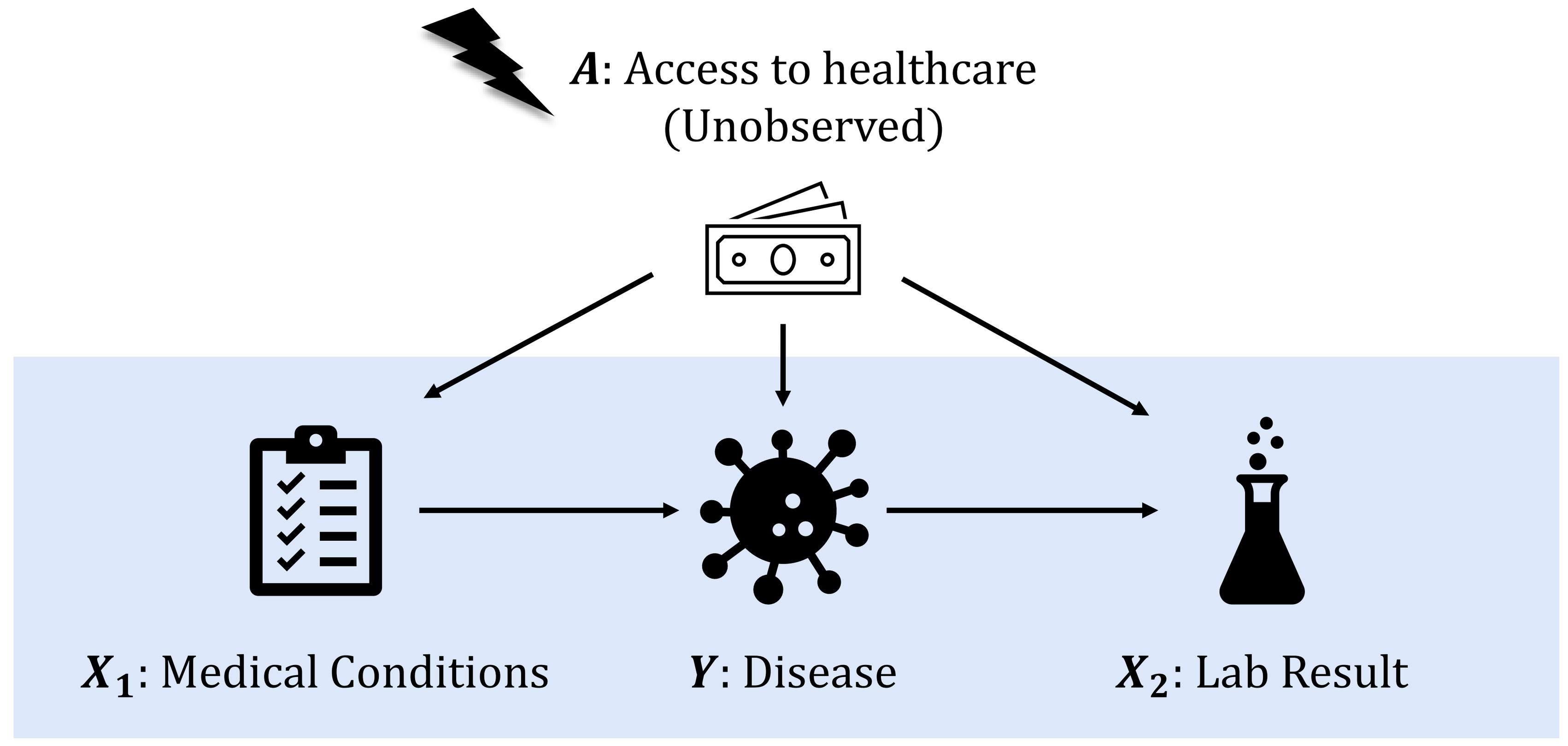
Regularizing towards Causal Invariance: Linear Models with Proxies

Mike Oberst MIT
Nikolaj Thams University of Copenhagen
Jonas Peters University of Copenhagen
David Sontag MIT

Motivation: Robustness to dataset shift

Challenge: Predictive performance may change due to changes in unobserved factors



Observed Distribution

Goal: Balance between in-distribution accuracy & robustness by minimizing a worst-case loss

$$\min_{\gamma} \sup_{v \in C} E_{do(A:=v)}[(Y - \gamma^\top X)^2]$$

Minimize worst-case loss over a set of interventions

Interventions on A change the distribution of P(X, Y)

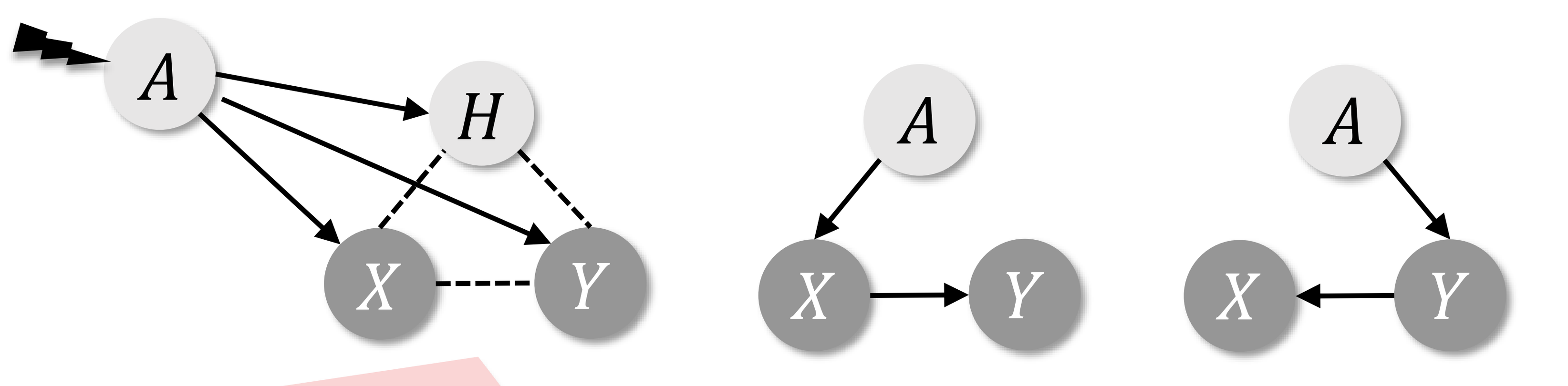
Previous Work: Anchor Regression [1] assumes that A is <u>observed</u> during training. **What can we do when A is not observed directly?**

Assumptions: Linear causal model + proxies

Linear Structural Causal Model (SCM) over all variables, and **noisy proxies of A**

A has no causal parents, but otherwise **any causal graph** over X, Y, H

Examples

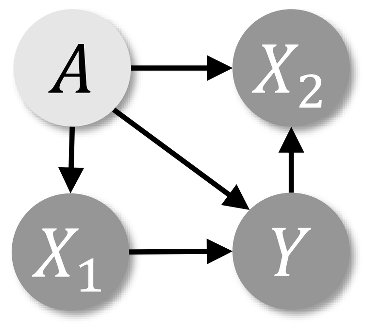


Unknown causal graph for X, Y, and hidden variables H.

Linear Structural Causal Model (SCM) over all variables

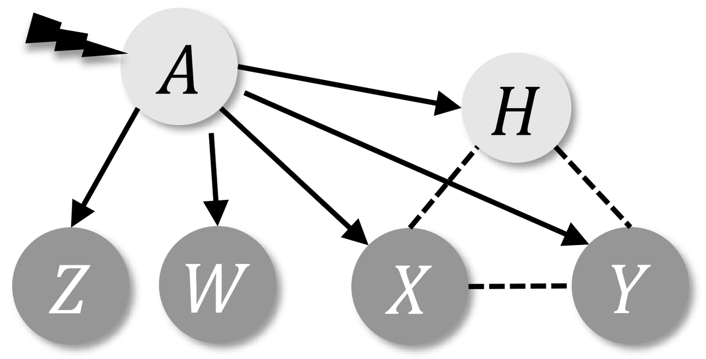
Example

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \epsilon$$



$$X_1 = \beta_1 A + \epsilon_{X_1}$$
$$Y = \beta_2 X_1 + \beta_3 A + \epsilon_Y$$
$$X_2 = \beta_4 Y + \beta_5 A + \epsilon_{X_2}$$

Proxies (available only during training) give independent views of A

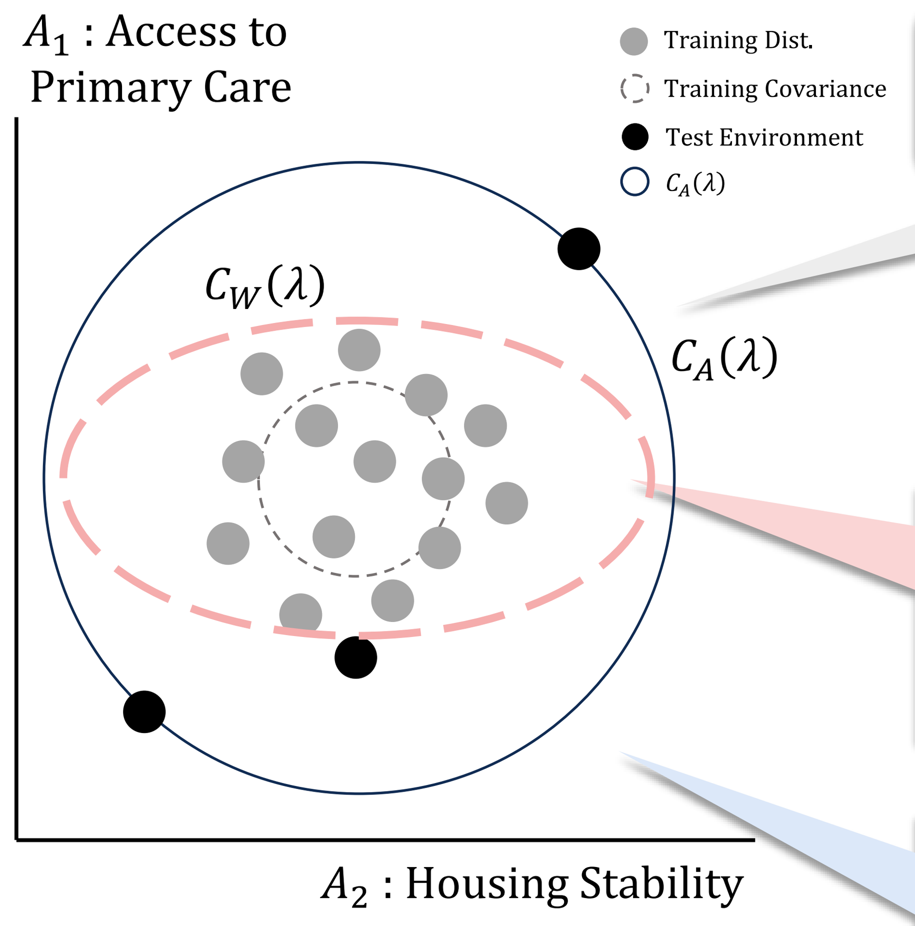


Noise terms are jointly independent

$$Z = \beta_Z^\top A + \epsilon_Z$$
$$W = \beta_W^\top A + \epsilon_W$$

Linear terms must be **full rank** to recover guarantees as if A were observed: Z, W, A must have the same dimension.

Recovering guarantees of Anchor Regression



X, Y not shown here, just the dimensions of A

Robustness set $C_A(\lambda)$ from Anchor Regression [1], when A is observed.

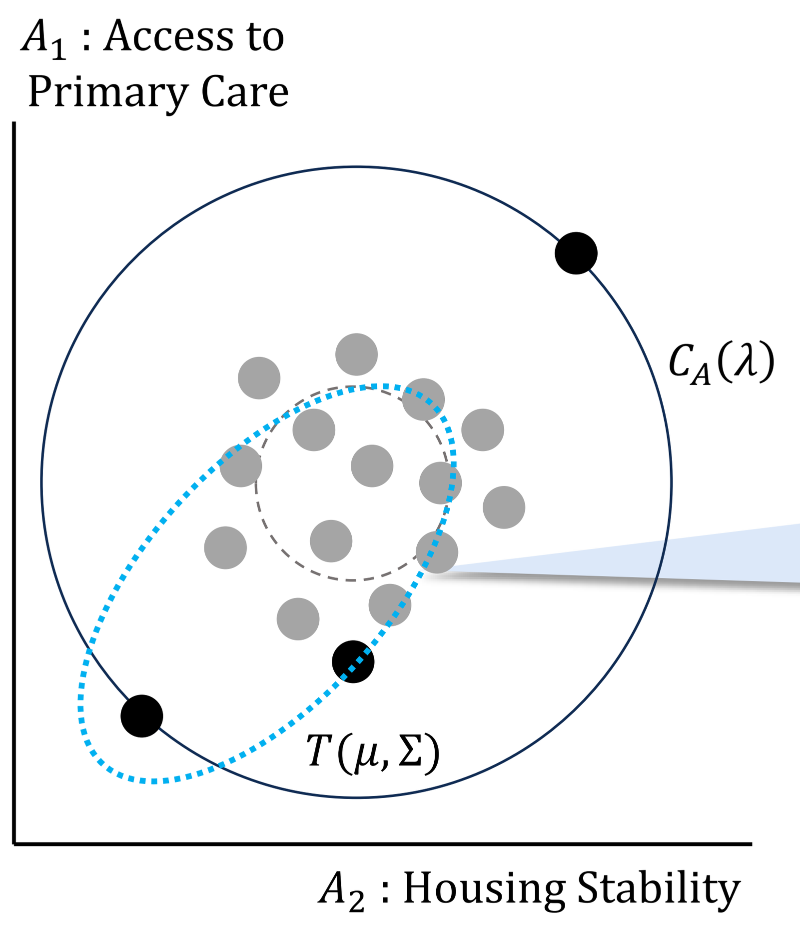
$$\sup_{v \in C_A(\lambda)} E_{do(A:=v)}[(Y - \gamma^\top X)^2]$$

$$C_A(\lambda) := \{v : E[vv^\top] \preccurlyeq (1+\lambda)\, E[AA^\top]\}$$

Theorem 1 (Informal)
Given a <u>single</u> noisy proxy W of A, the robustness set is provably reduced, and this reduction is not identifiable

Theorem 2 (Informal)
Given <u>two</u> noisy proxies of A, one can recover the original robustness set, using a <u>modified objective</u>

Targeting the robustness set to anticipated shifts



Motivation: What if we have some prior knowledge? For instance, moving to a hospital that serves a population with less access to healthcare & housing?

Theorem 3 (Informal)
We generalize to a larger class of robustness sets, and prove that the minimizer is identified with <u>two proxies</u>

$$\sup_{v \in T(\mu,\Sigma)} E_{do(A:=v)}[(Y - \gamma^\top X)^2]$$

$$T(\mu, \Sigma) := \{v : E[(v-\mu)(v-\mu)^\top] \preccurlyeq \Sigma\}$$

Code available at github.com/clinicalml/proxy-anchor-regression

[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.