Regularizing towards Causal Invariance: Linear Models with Proxies


Mike Oberst
MIT

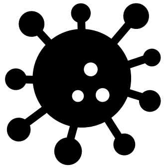

Nikolaj Thams
Univ. of Copenhagen


Jonas Peters
Univ. of Copenhagen


David Sontag
MIT

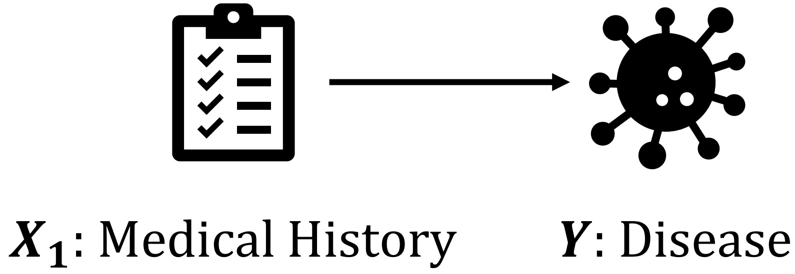
@MichaelOberst

Motivation: Robustness to Dataset Shift

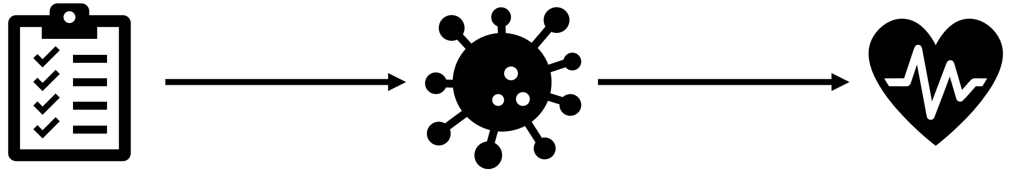


$\boldsymbol{Y}$: Disease

Motivation: Robustness to Dataset Shift

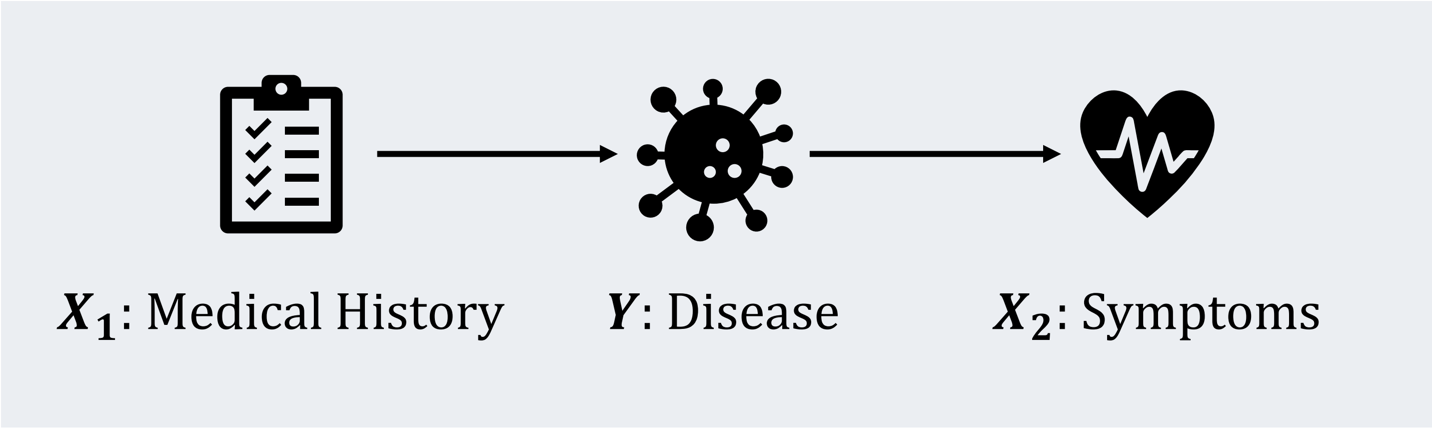


Motivation: Robustness to Dataset Shift



X_1: Medical History Y: Disease X_2: Symptoms

Motivation: Robustness to Dataset Shift

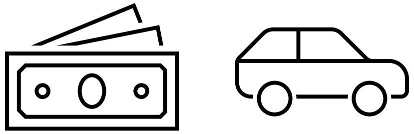


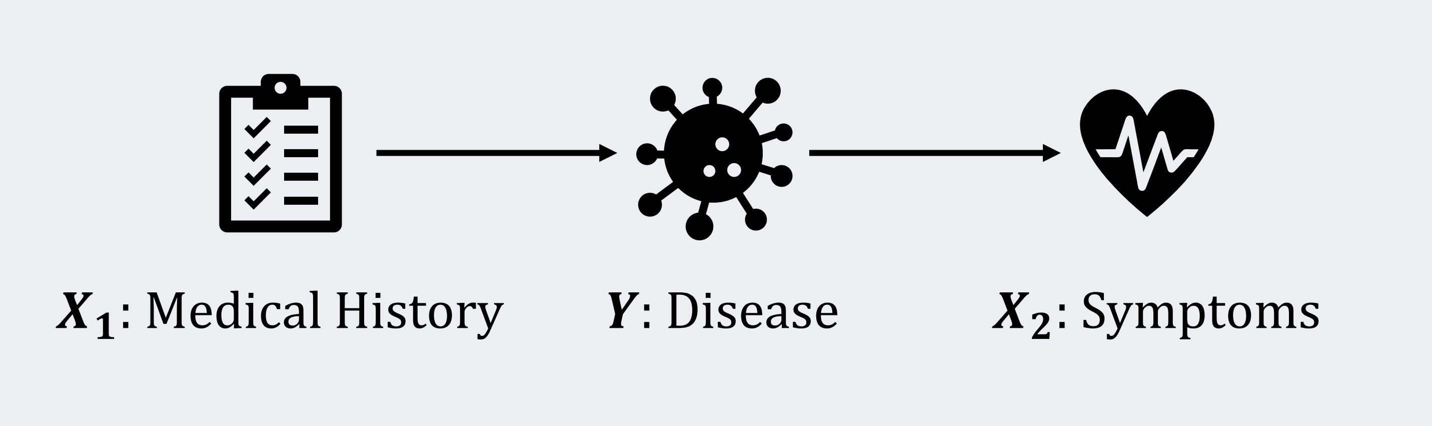
Observed Distribution

Motivation: Robustness to Dataset Shift

Examples: Income, distance to nearest clinic.

$\boldsymbol{A}$: Access to healthcare (Unobserved)



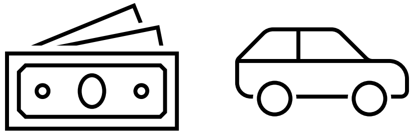


Observed Distribution

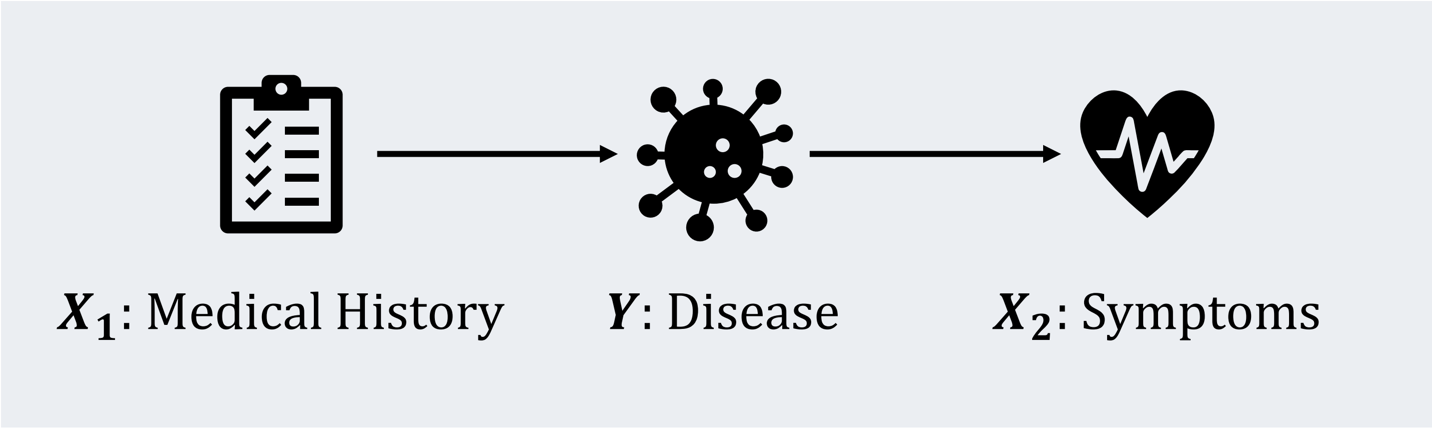
Motivation: Robustness to Dataset Shift

Examples: Income, distance to nearest clinic.

A: Access to healthcare (Unobserved)

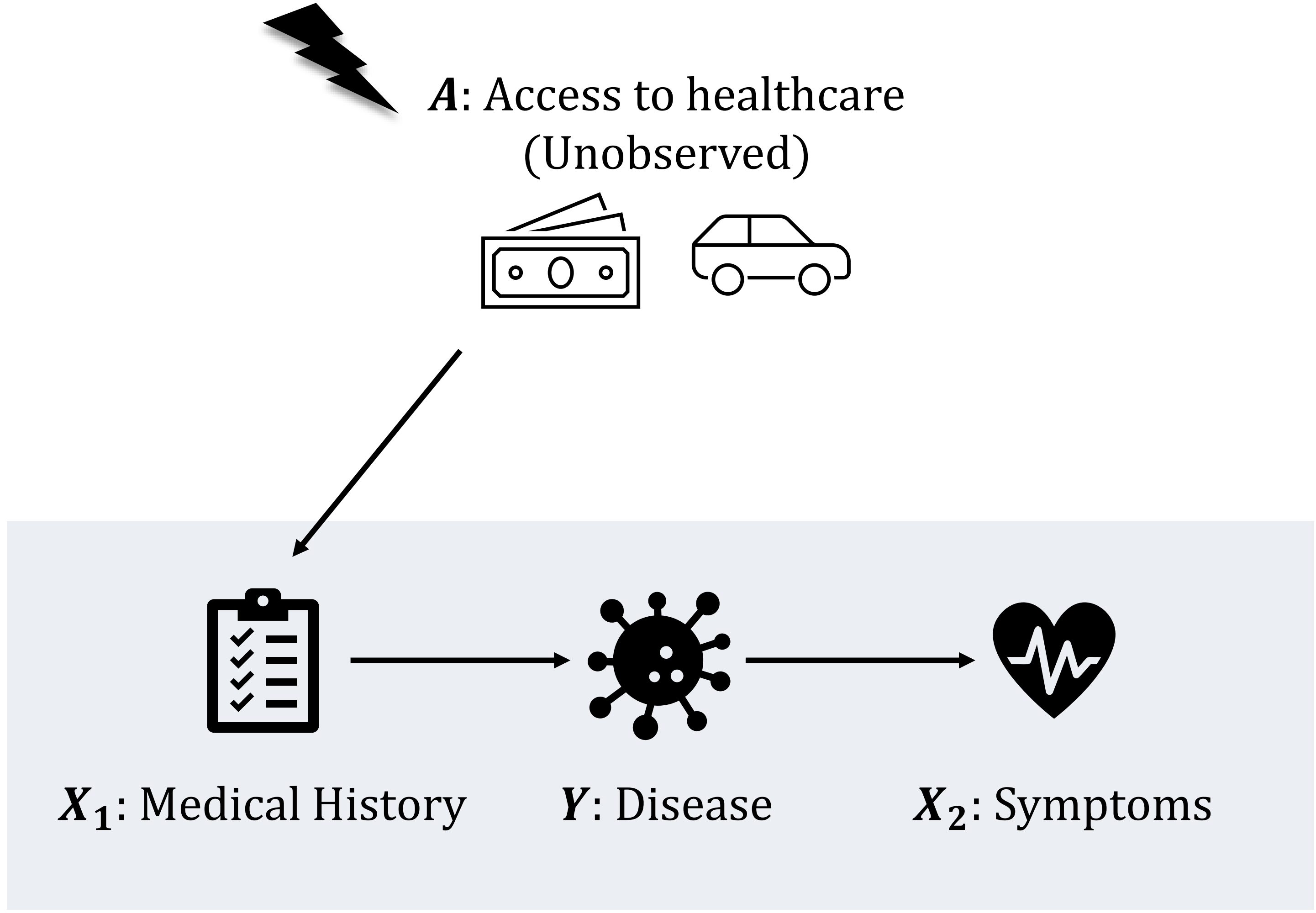


How do changes here (e.g., due to differences between hospitals) influence our predictive models?



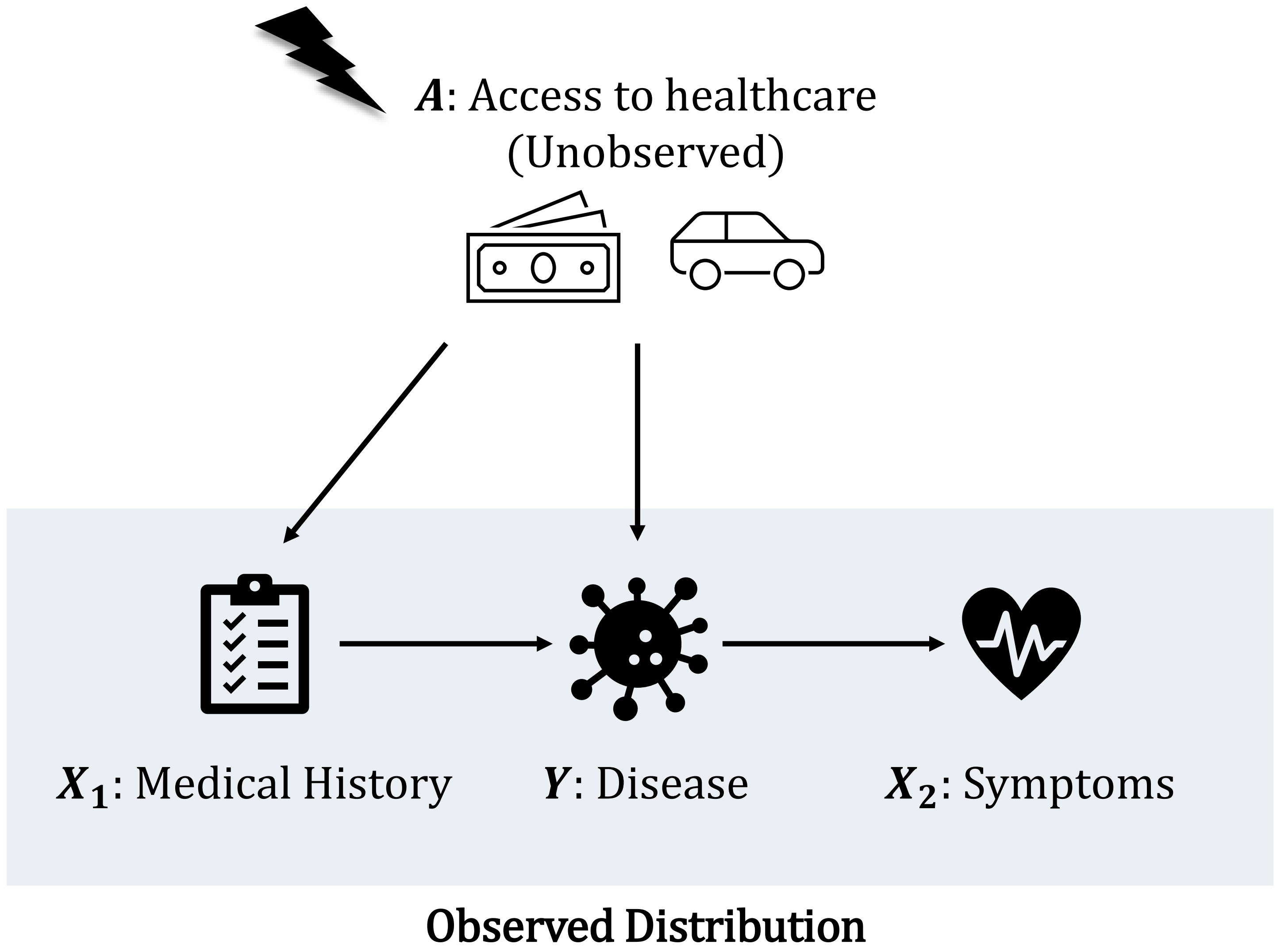
Observed Distribution

Motivation: Robustness to Dataset Shift



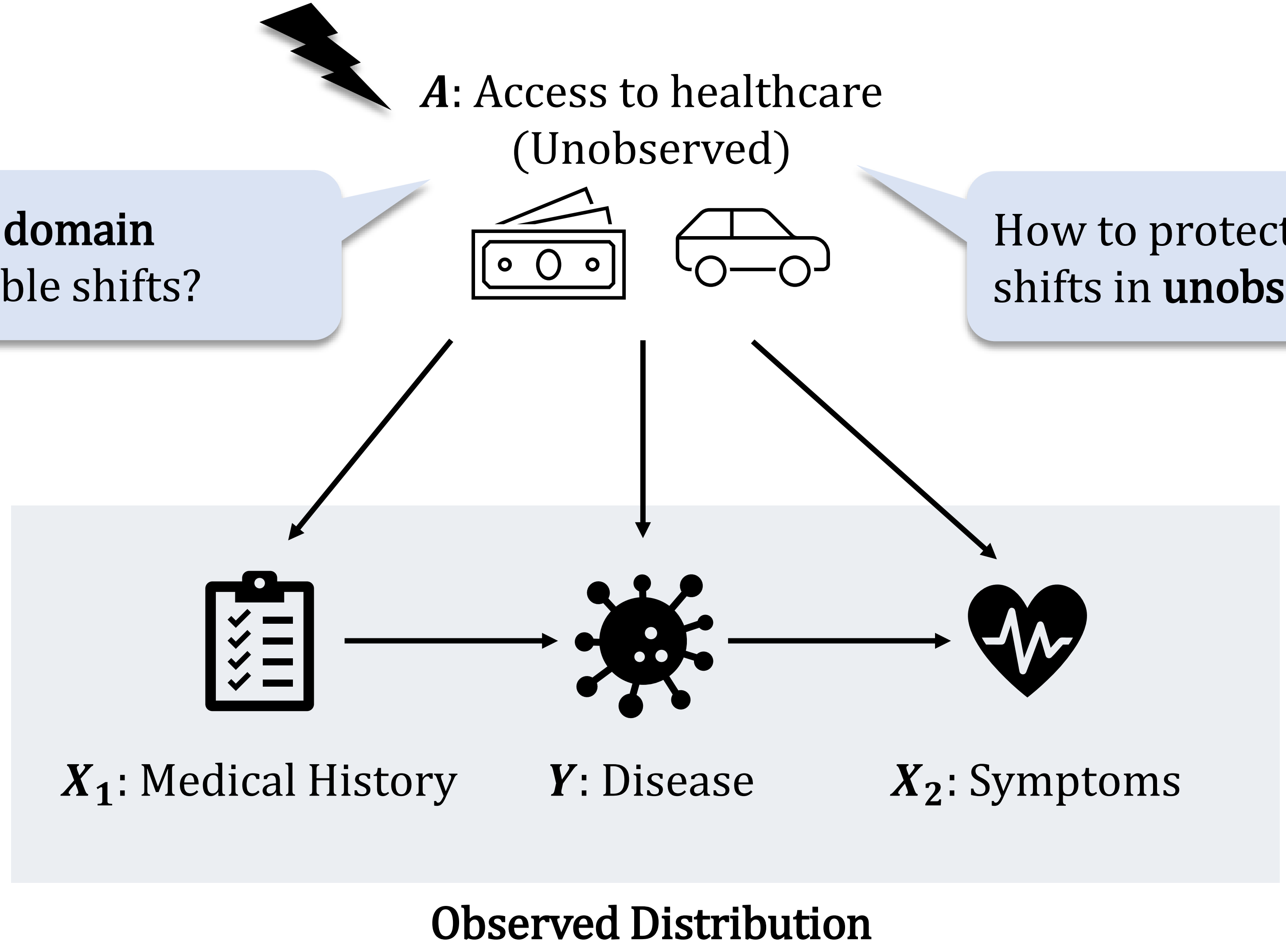
Observed Distribution

Motivation: Robustness to Dataset Shift



Motivation: Robustness to Dataset Shift

How to incorporate **domain knowledge** of plausible shifts?



How to protect against plausible shifts in **unobserved** factors?

Our contributions

Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

High-Level Overview

- Setup: Linear SCMs and Shift Interventions
- Background: Robustness to bounded shift in linear models
- Contributions:
  - Defining (and optimizing over) more flexible robustness sets
  - Recovering guarantees with noisy proxies

High-Level Overview

- Setup: Linear SCMs and Shift Interventions
- Background: Robustness to bounded shift in linear models
- Contributions:
  - Defining (and optimizing over) more flexible robustness sets
  - Recovering guarantees with noisy proxies

Setup: Linear SCMs and Shift Interventions

Assumptions: Linear structural causal model (SCM) over all observed and unobserved variables and one or more **noisy proxies of A**

Setup: Linear SCMs and Shift Interventions

Assumptions: Linear structural causal model (SCM) over all observed and unobserved variables and one or more **noisy proxies of A**

Any causal graph over X, Y, H is permitted, but A is an “anchor” with no causal parents.

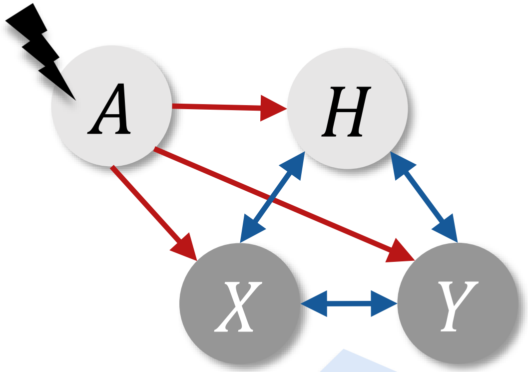
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \epsilon$$

[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.

Setup: Linear SCMs and Shift Interventions

Assumptions: Linear structural causal model (SCM) over all observed and unobserved variables and one or more **noisy proxies of A**

Any causal graph over X, Y, H is permitted, but A is an "anchor" with no causal parents.



Unknown structure on (X, Y, H)

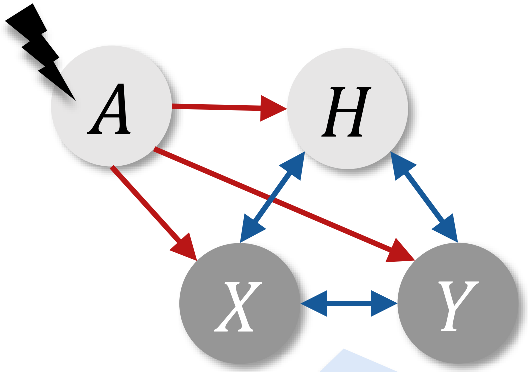
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \epsilon$$

Setup: Linear SCMs and Shift Interventions

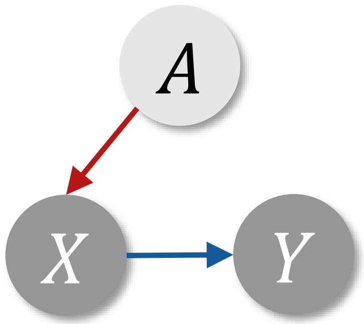
Assumptions: Linear structural causal model (SCM) over all observed and unobserved variables and one or more **noisy proxies of A**

Any causal graph over X, Y, H is permitted, but A is an “anchor” with no causal parents.

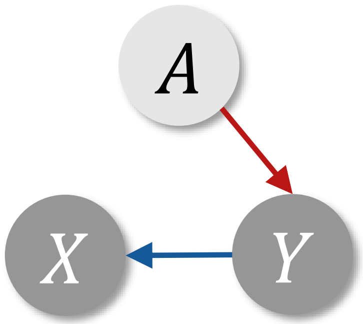
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + MA + \epsilon$$



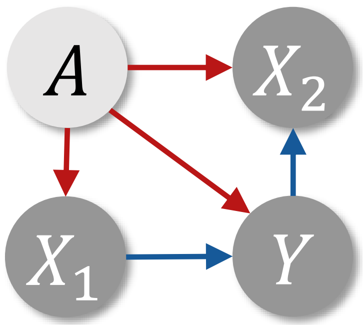
Unknown structure on (X, Y, H)



Covariate Shift



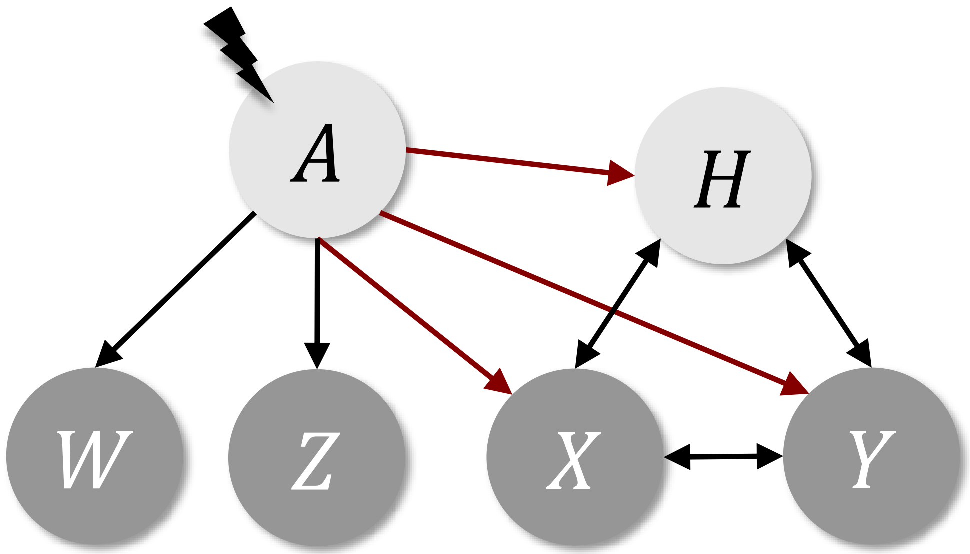
Label Shift



...and more

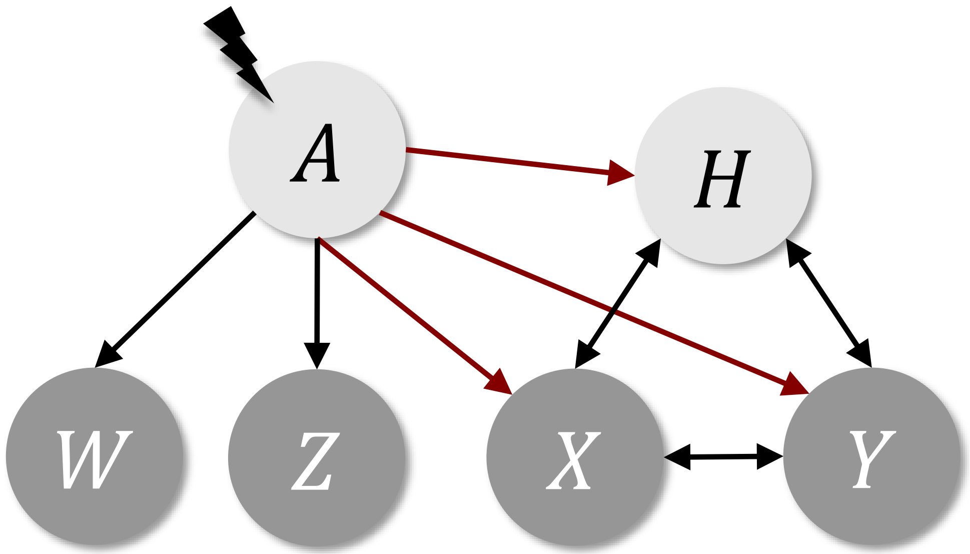
Goal: Robustness to Dataset Shift

Use noisy proxies (W, Z), only available at training time…



Goal: Robustness to Dataset Shift

Use noisy proxies (W, Z), only available at training time…



… to learn a model that **minimizes a worst-case loss** over interventions on A

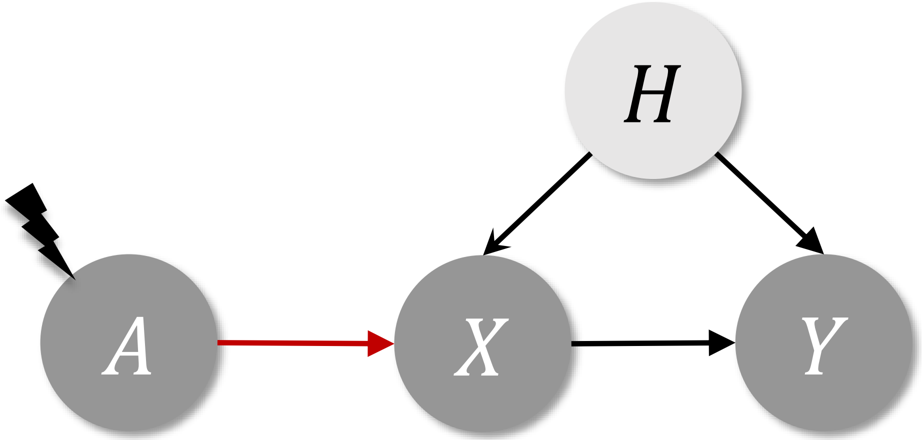
$$\min \sup_{\nu \in C} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

High-Level Overview

- Setup: Linear SCMs and Shift Interventions
- Background: Robustness to bounded shift in linear models
- Contributions:
  - Defining (and optimizing over) more flexible robustness sets
  - Recovering guarantees with noisy proxies

Robustness to bounded interventions

Example: Intervention on Instrumental Variable (IV)



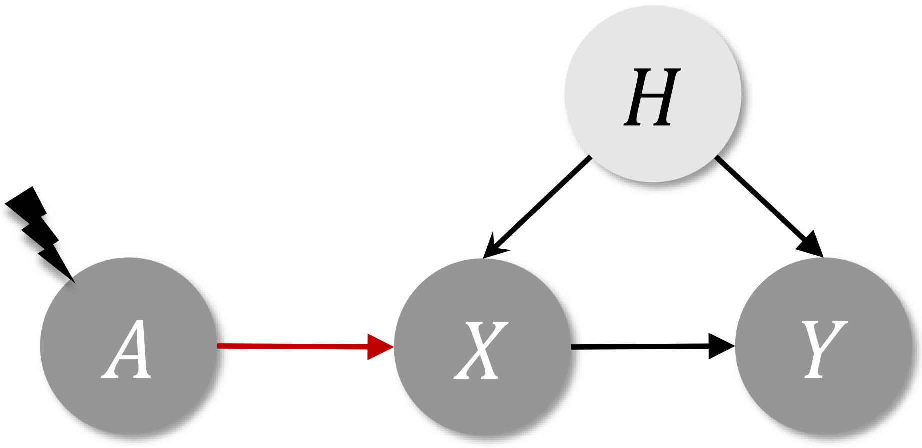
Linear functions with additive noise.

$X = A + H + \epsilon_X \qquad Y = X + 2H + \epsilon_Y$

Modified version of Example 2 from 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.

Robustness to bounded interventions

Example: Intervention on Instrumental Variable (IV)



Linear functions with additive noise.

$X = A + H + \epsilon_X \qquad Y = X + 2H + \epsilon_Y$

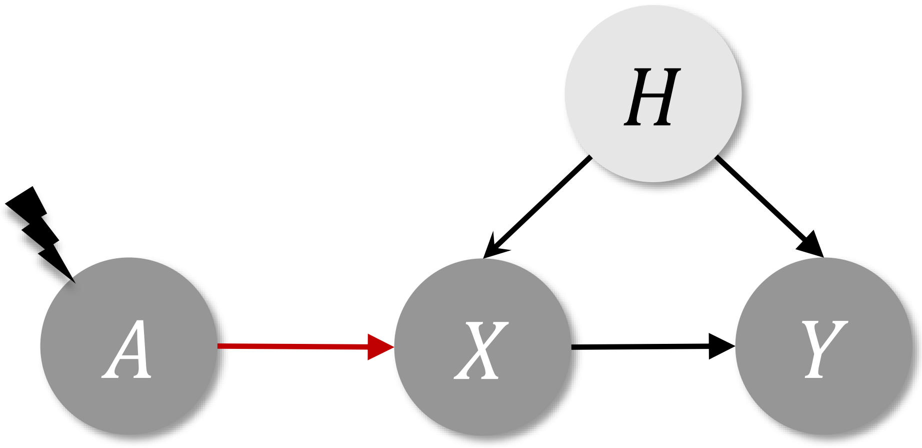
Which linear predictor $(\gamma \cdot X)$ to use?

$$\gamma_{causal} = 1, \ \gamma_{OLS} \neq 1$$

Modified version of Example 2 from 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.

Robustness to bounded interventions

Example: Intervention on Instrumental Variable (IV)



Linear functions with additive noise.

$X = A + H + \epsilon_X \qquad Y = X + 2H + \epsilon_Y$

Which linear predictor $(\gamma \cdot X)$ to use?

$$\gamma_{causal} = 1, \ \gamma_{OLS} \neq 1$$

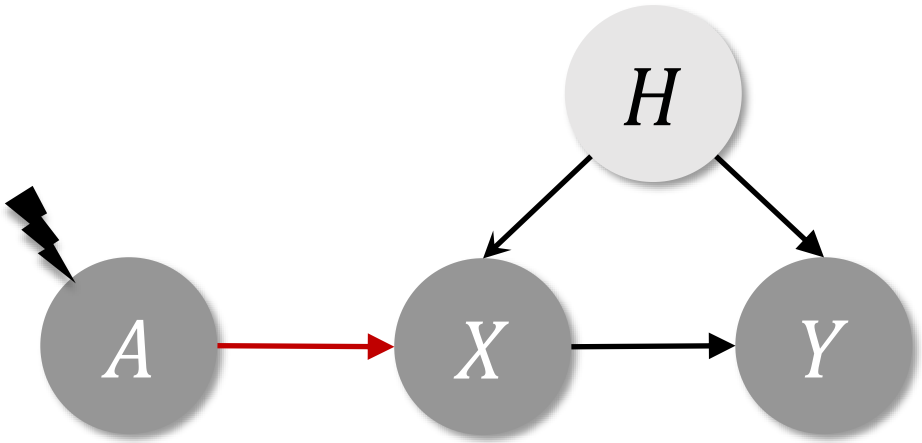
Note: X, Y *are both linear functions of* A*:*

$$Y = (A + H + \epsilon_X) + 2H + \epsilon_Y$$

Modified version of Example 2 from 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.

Robustness to bounded interventions

Example: Intervention on Instrumental Variable (IV)



Linear functions with additive noise.

$X = A + H + \epsilon_X \qquad Y = X + 2H + \epsilon_Y$

Which linear predictor ($\gamma \cdot X$) to use?

$$\gamma_{causal} = 1, \; \gamma_{OLS} \neq 1$$

Residual is a linear function of A...

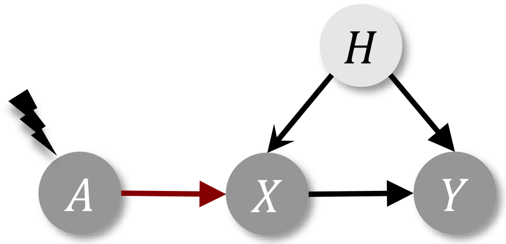
$$Y - \gamma \cdot X = (1 - \gamma)A + \cdots$$

Remainder does not depend on A

Causal effect yields **invariant** performance under **arbitrary** interventions on A

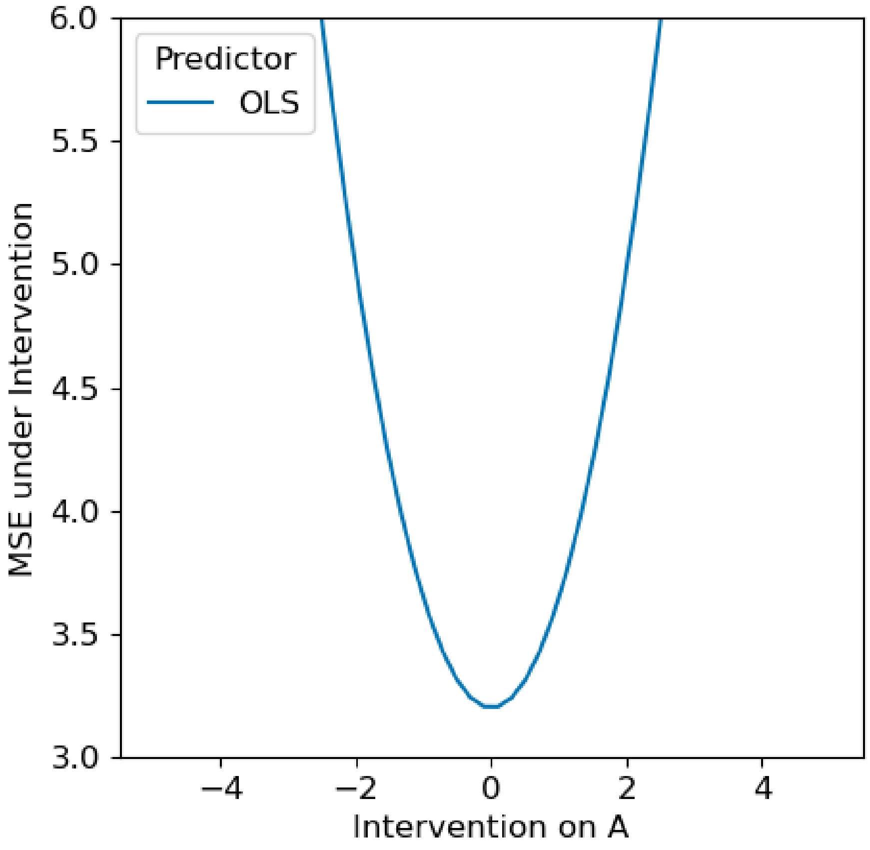
Modified version of Example 2 from Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Robustness to bounded interventions



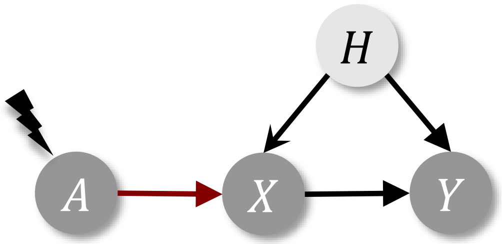
Plot: MSE of different predictors under interventions on A

- Ordinary least-squares (OLS)



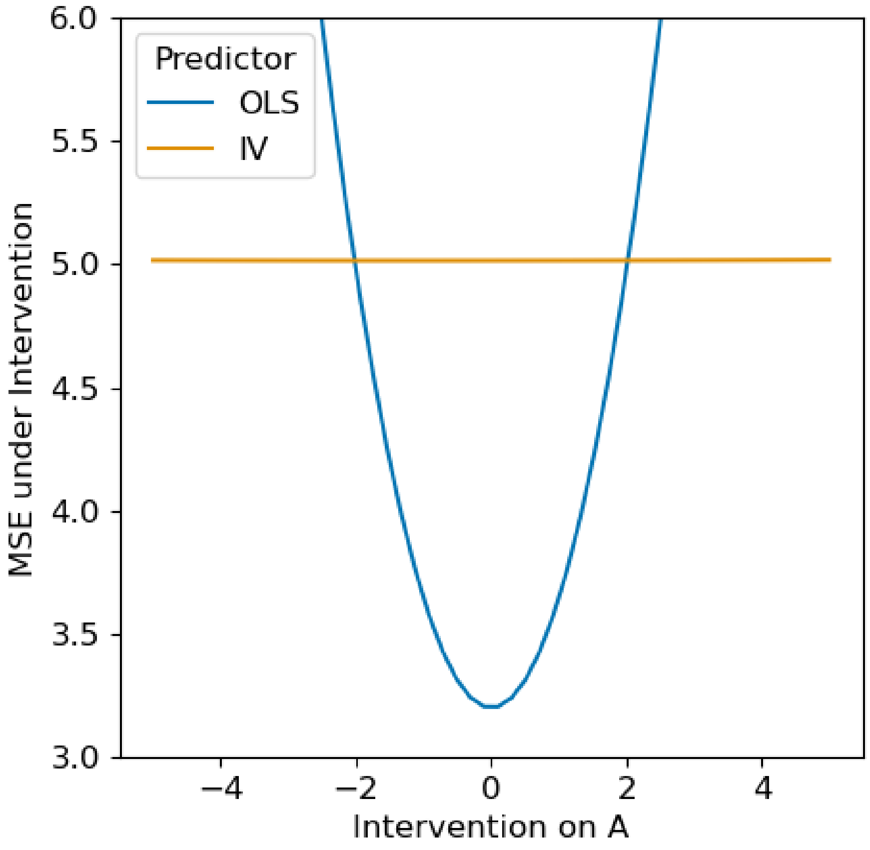
Intervention: Set A to a fixed value (similar plot holds for shift in the mean of A)

Robustness to bounded interventions



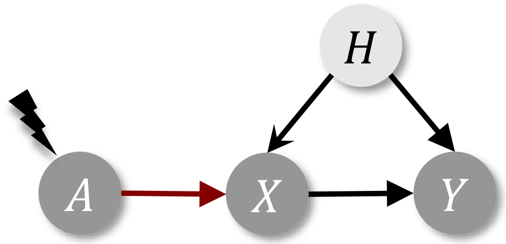
Plot: MSE of different predictors under interventions on A

- Ordinary least-squares (OLS)
- Causal effect (IV)



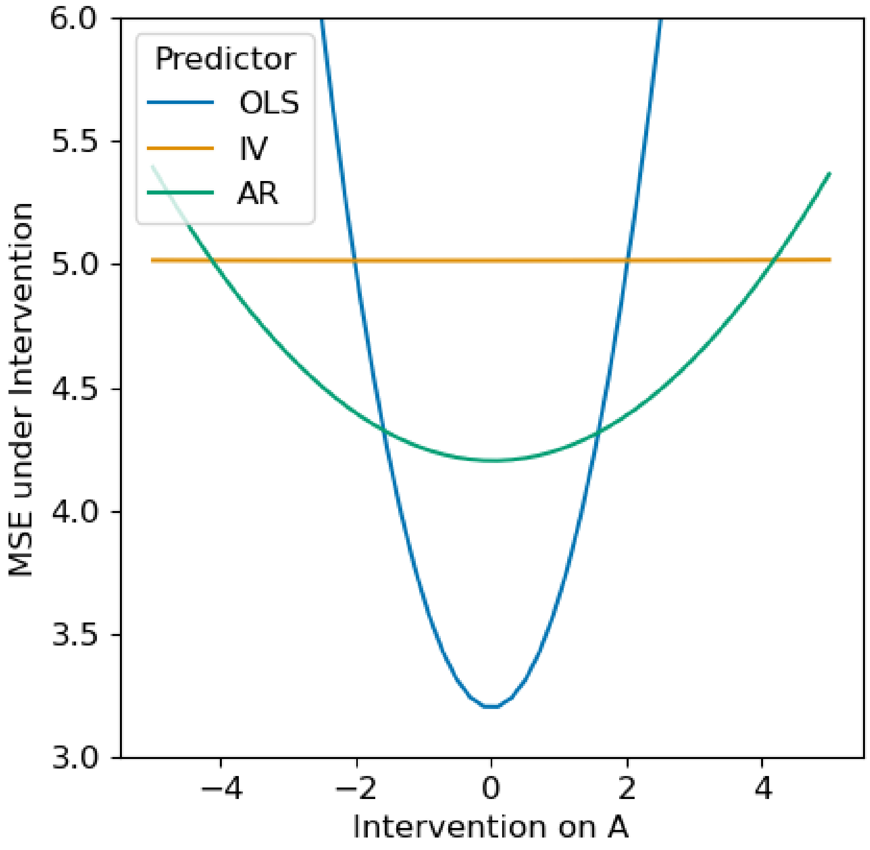
Intervention: Set A to a fixed value (similar plot holds for shift in the mean of A)

Robustness to bounded interventions



Plot: MSE of different predictors under interventions on A

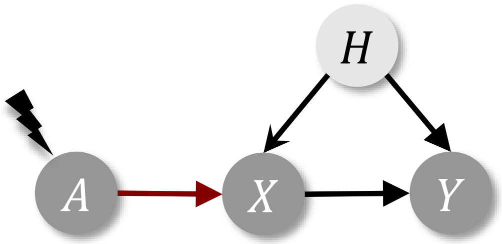
- Ordinary least-squares (OLS)
- Causal effect (IV)
- Anchor Regression (AR, $\lambda = 6$)



Intervention: Set A to a fixed value (similar plot holds for shift in the mean of A)

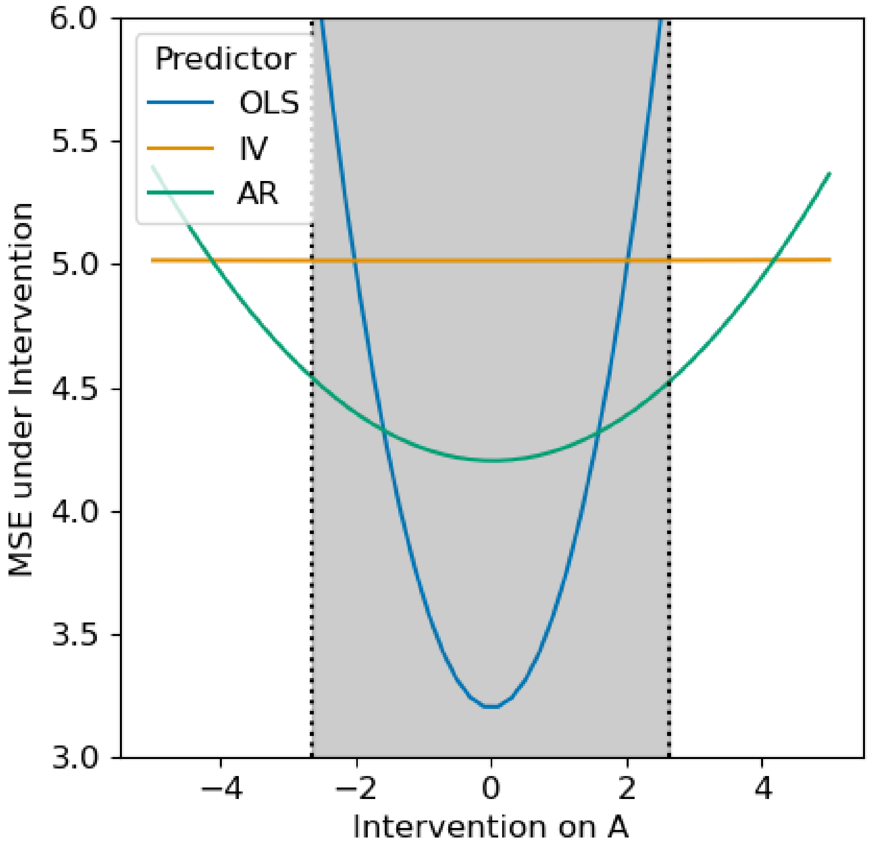
[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Robustness to bounded interventions



Plot: MSE of different predictors under interventions on A

- Ordinary least-squares (OLS)
- Causal effect (IV)
- Anchor Regression (AR, $\lambda = 6$)



Note: Anchor Regression has optimal worst-case risk for interventions in this range [1]

[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.

$$R(\gamma) \coloneqq Y - \gamma^\top X$$

Anchor Regression

Idea: Trade off between invariance & in-distribution accuracy

$$\ell_{AR}(X, Y, A; \gamma, \lambda) = \underbrace{\ell_{LS}(X, Y; \gamma)}_{\mathbb{E}[R(\gamma)^2]} + \lambda \cdot \underbrace{\ell_{PLS}(X, Y, A; \gamma)}_{\mathbb{E}[(\mathbb{E}[\,R(\gamma) \mid A\,])^2]}$$

[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.

$$R(\gamma) \coloneqq Y - \gamma^\top X$$

Anchor Regression

Least Squares Loss

Penalize influence of A on the residual

$$\ell_{AR}(X,Y,A;\gamma,\lambda) = \underbrace{\ell_{LS}(X,Y;\gamma)}_{\mathbb{E}[R(\gamma)^2]} + \lambda \cdot \underbrace{\ell_{PLS}(X,Y,A;\gamma)}_{\mathbb{E}[(\mathbb{E}[\,R(\gamma) \mid A\,])^2]}$$

Theorem 1 (Rothenhausler 2021)

$$\ell_{AR}(A;\gamma,\lambda) = \sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A \coloneqq \nu)}[R(\gamma)^2]$$

$$C_A(\lambda) \coloneqq \{\nu : \nu\nu^\top \preccurlyeq (1+\lambda)\,\mathbb{E}[AA^\top]\,\}$$

[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.

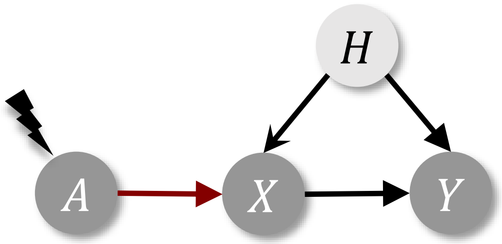
High-Level Overview

- Setup: Linear SCMs and Shift Interventions
- Background: Robustness to bounded shift in linear models
- Contributions:
  - Defining (and optimizing over) more flexible robustness sets
  - Recovering guarantees with noisy proxies

Our contributions

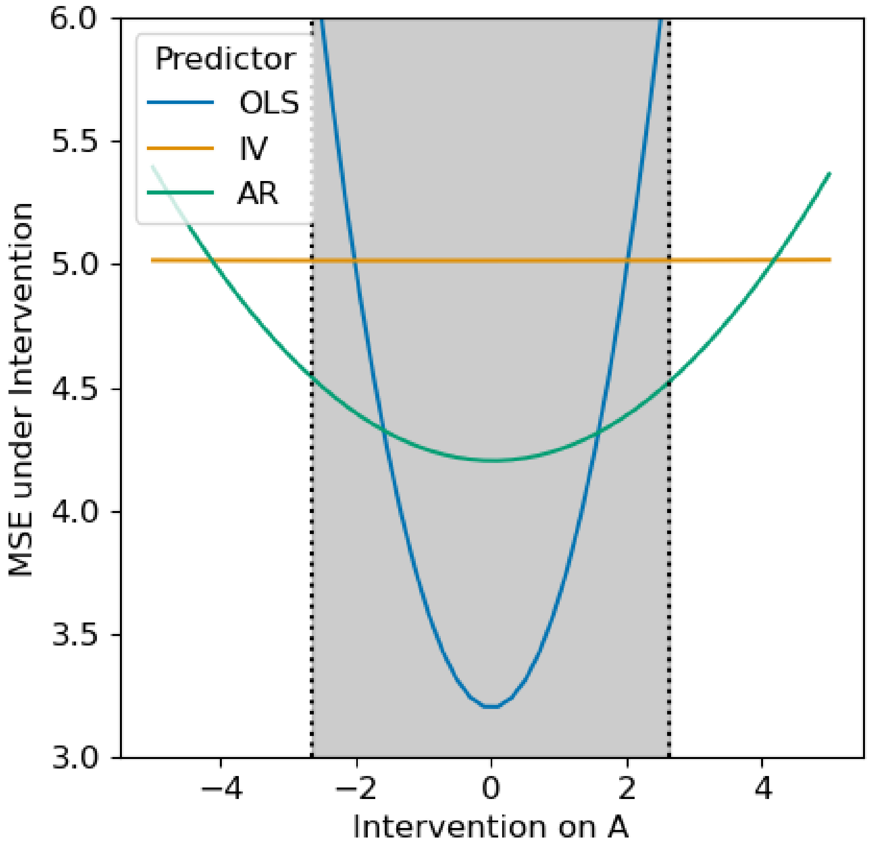
Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

Robustness to targeted interventions

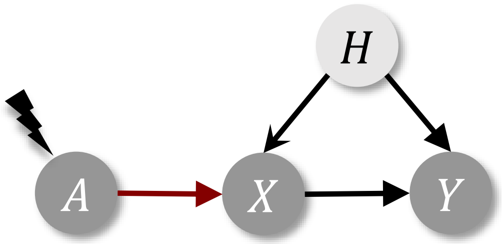


Problem: What if we have more specific knowledge of the shift?

E.g., moving to a hospital with a ***lower*** *level of income, but not higher.*

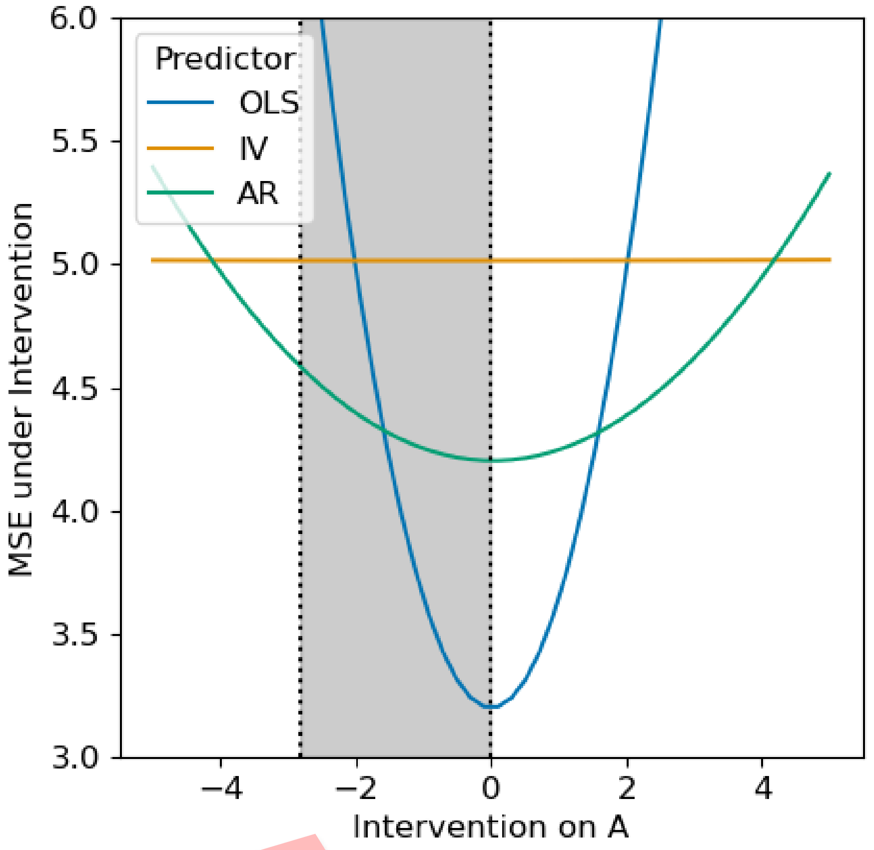


Robustness to targeted interventions



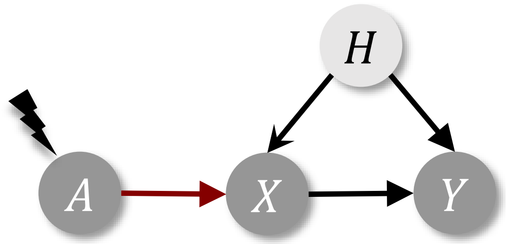
Problem: What if we have more specific knowledge of the shift?

*E.g., moving to a hospital with a **lower** level of income, but not higher.*



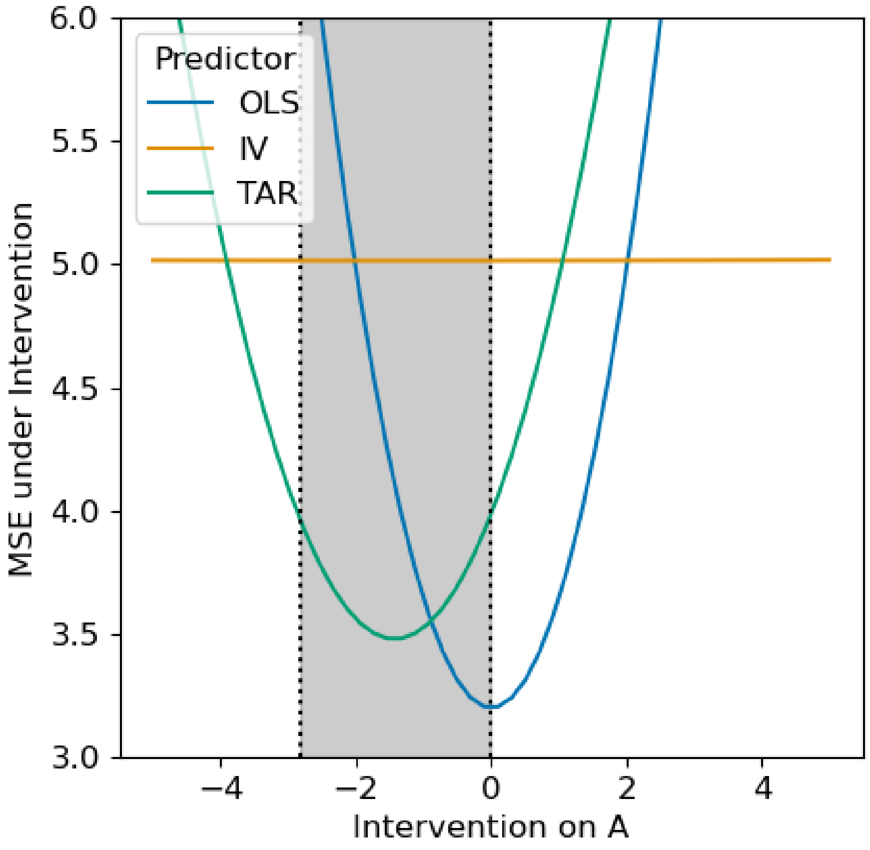
How to protect against this shift?

Robustness to targeted interventions



Problem: What if we have more specific knowledge of the shift?

*E.g., moving to a hospital with a **lower** level of income, but not higher.*

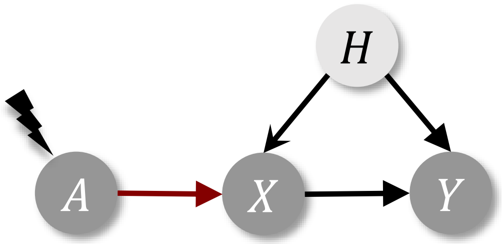


Contribution: We show how to adapt the guarantees of Anchor Regression to a broader class of robustness sets

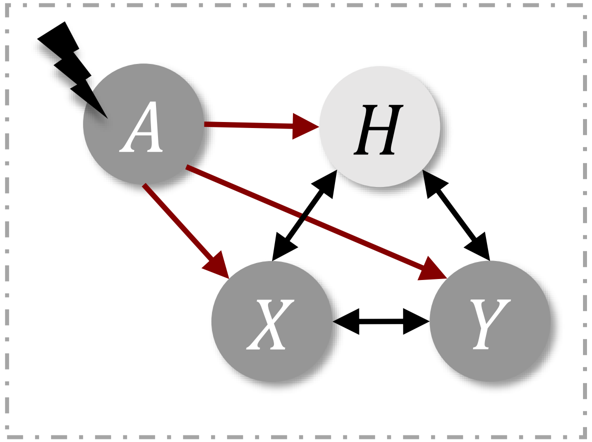
Our contributions

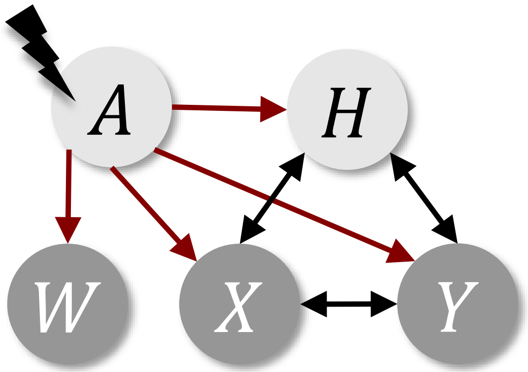
Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

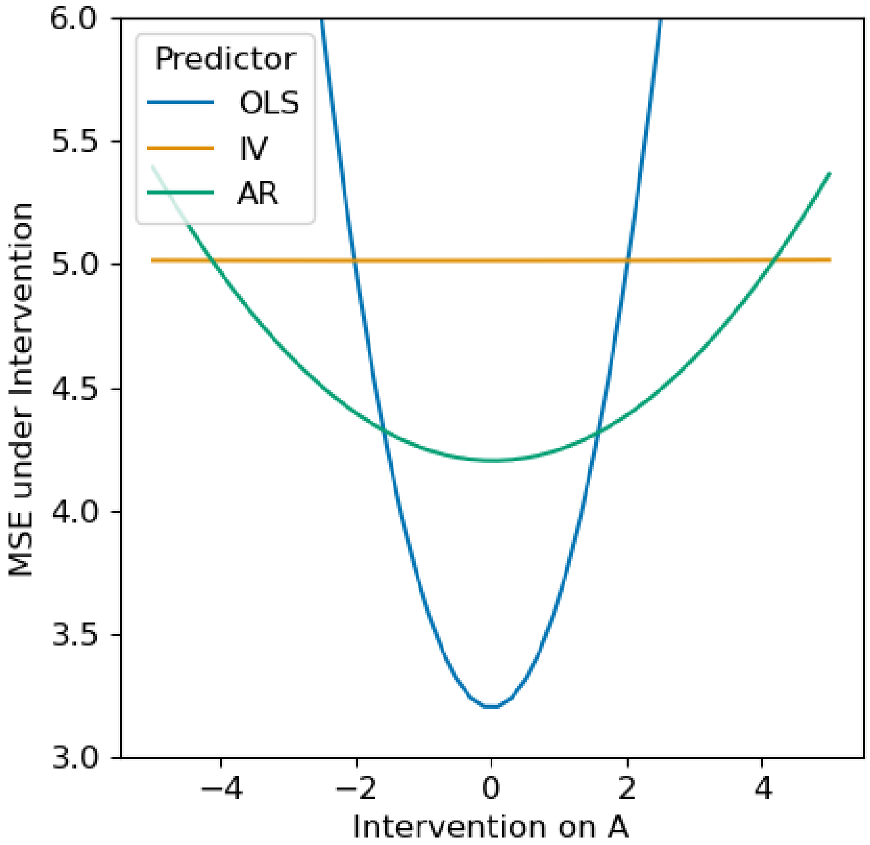
Robustness with noisy proxies



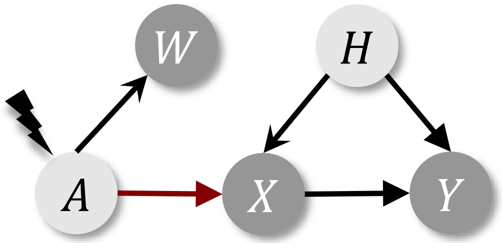
Problem: What happens when A is only observed with noise?



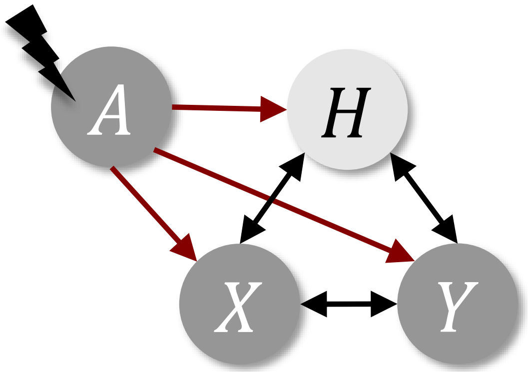


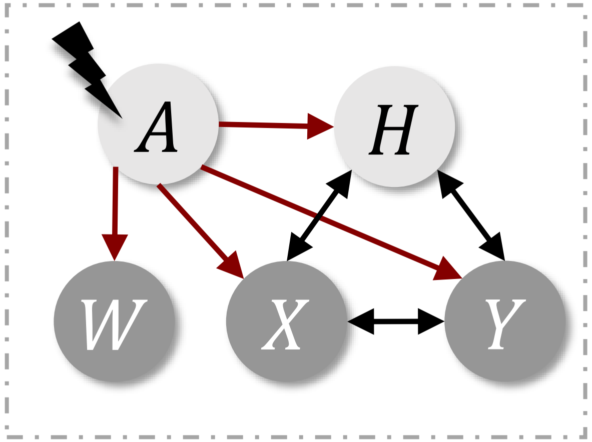


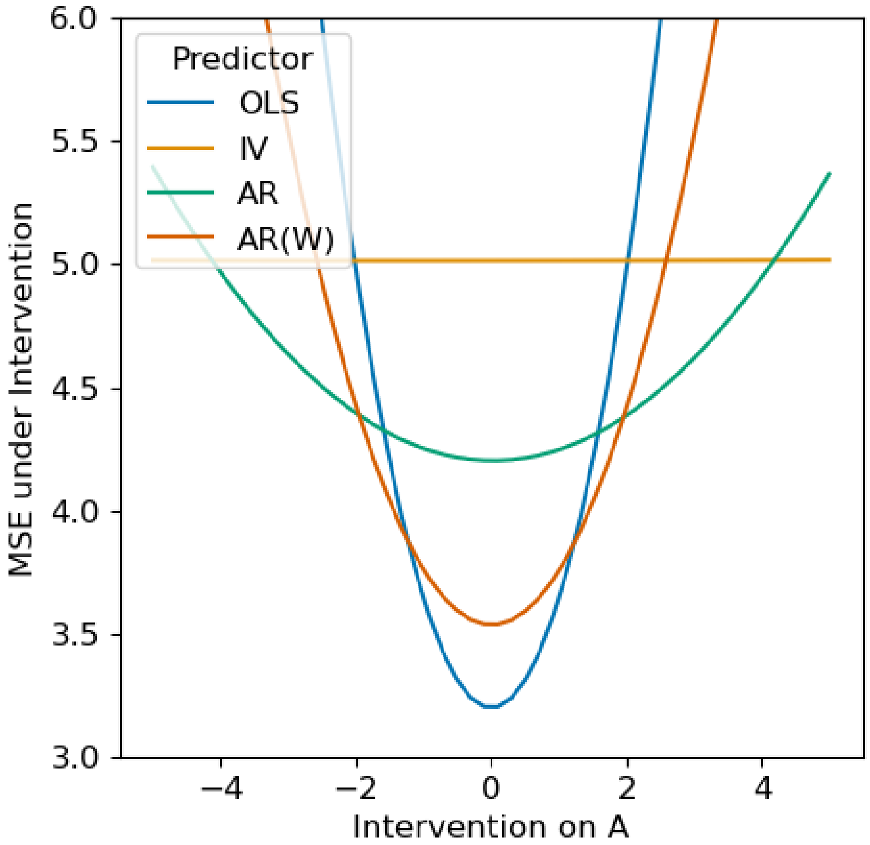
Robustness with noisy proxies



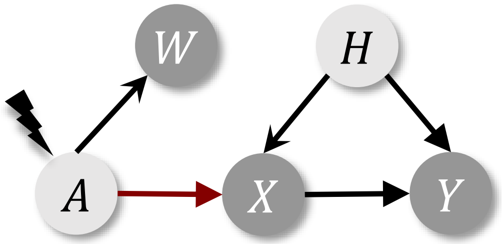
Problem: What happens when A is only observed with noise?



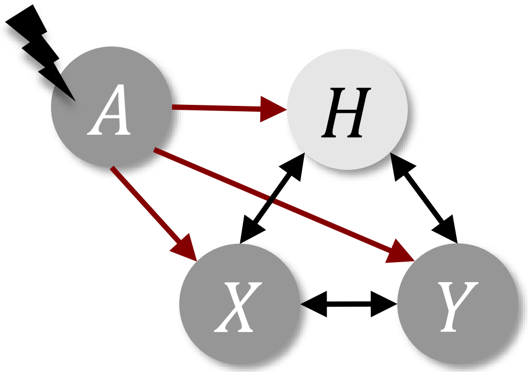


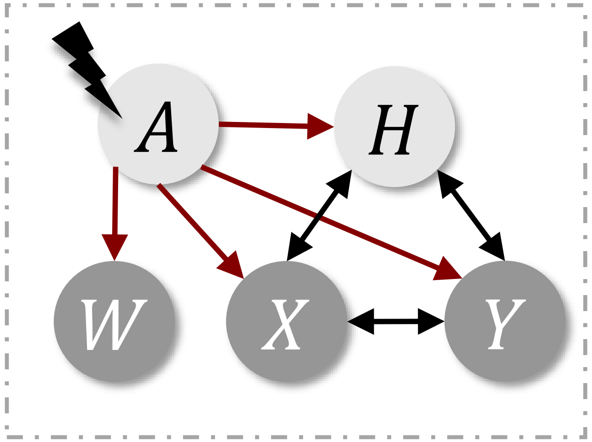


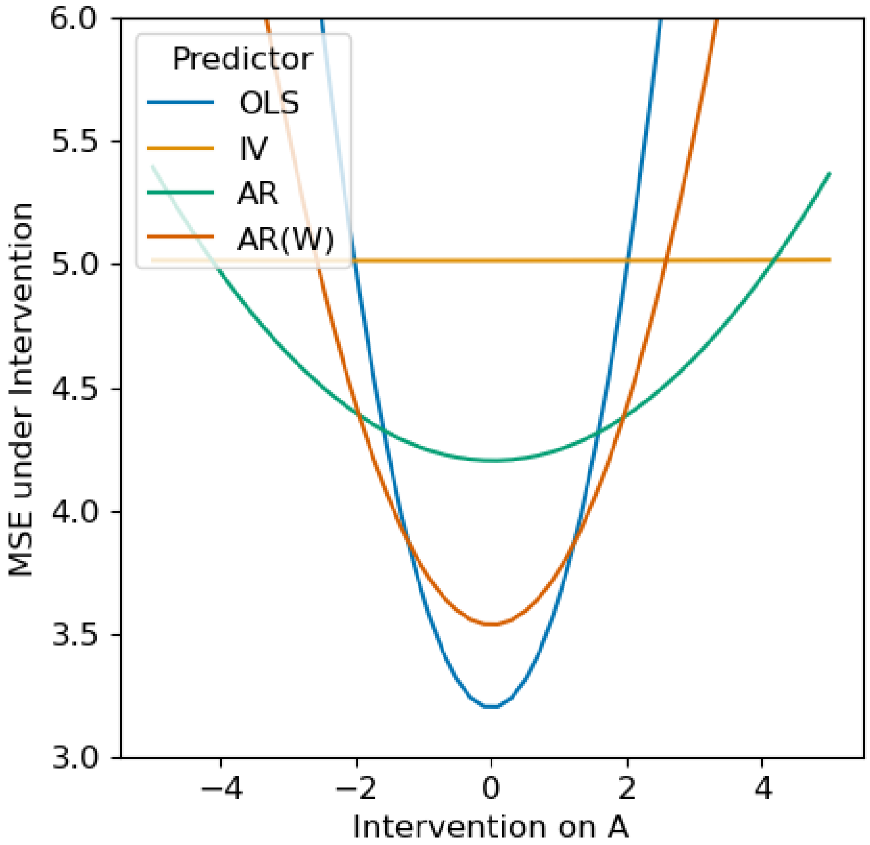
Robustness with noisy proxies



Problem: What happens when A is only observed with noise?

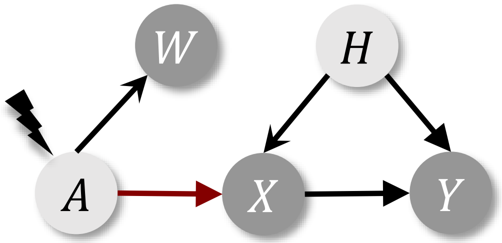




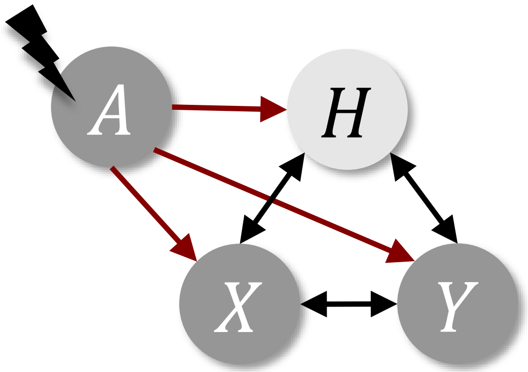


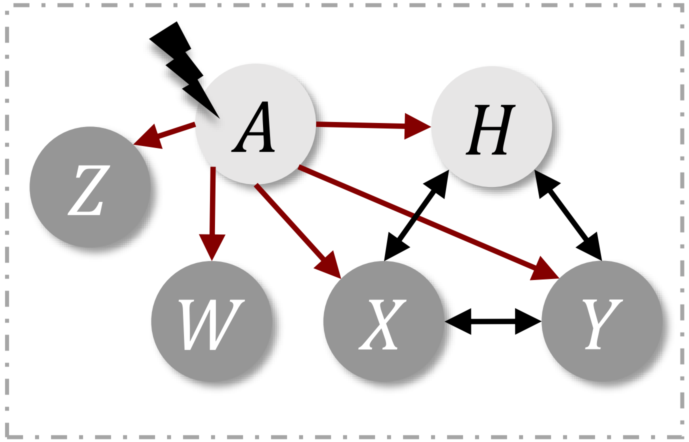
Contribution: We demonstrate that proxy noise reduces robustness with a single proxy

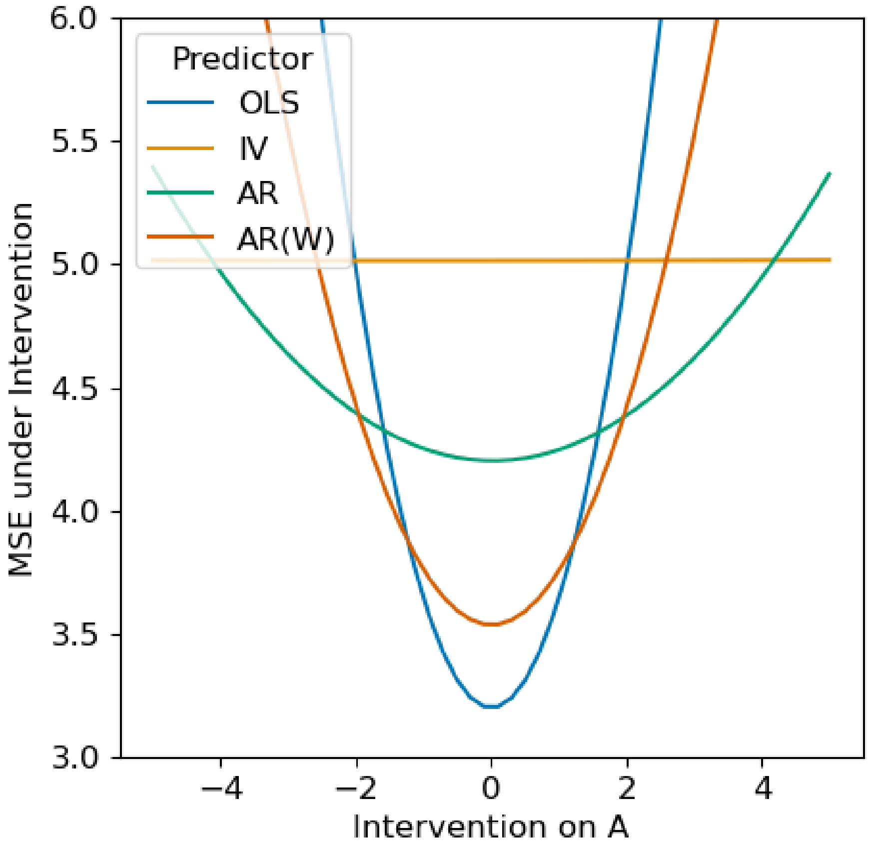
Robustness with noisy proxies



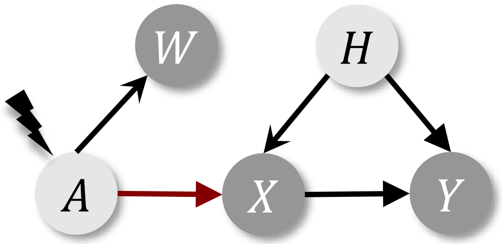
Problem: What happens when A is only observed with noise?



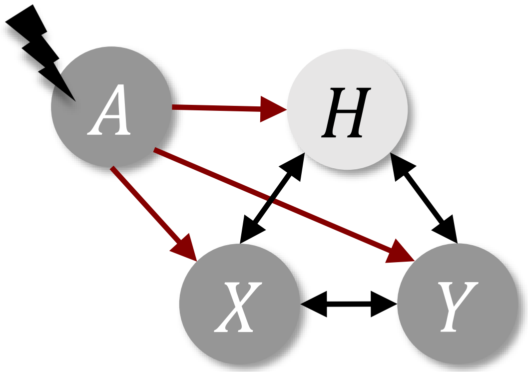


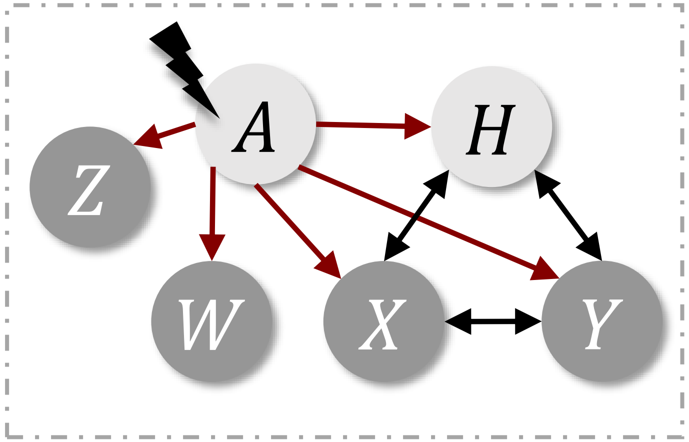


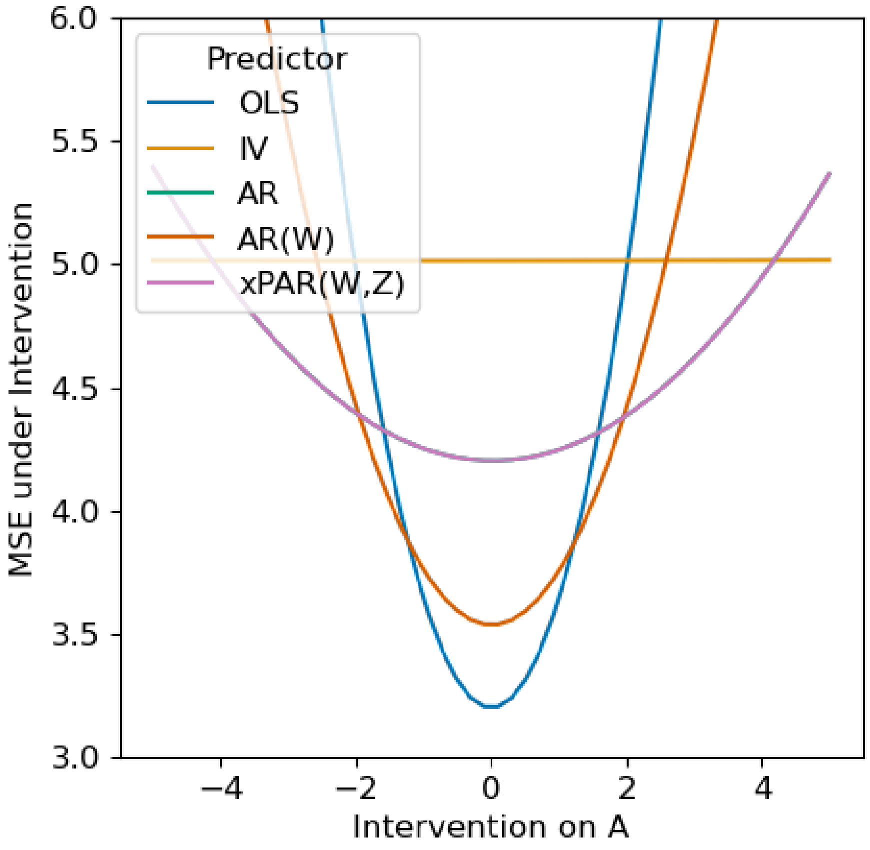
Robustness with noisy proxies



Problem: What happens when A is only observed with noise?







Contribution: We demonstrate that two proxies can be used to recover the original guarantees, as if A were observed

Our contributions

1

Optimize worst-case loss over interventions on A in a targeted robustness set.

Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

2

Two proxies suffice to recover guarantees as if A were observed.

Our contributions

1

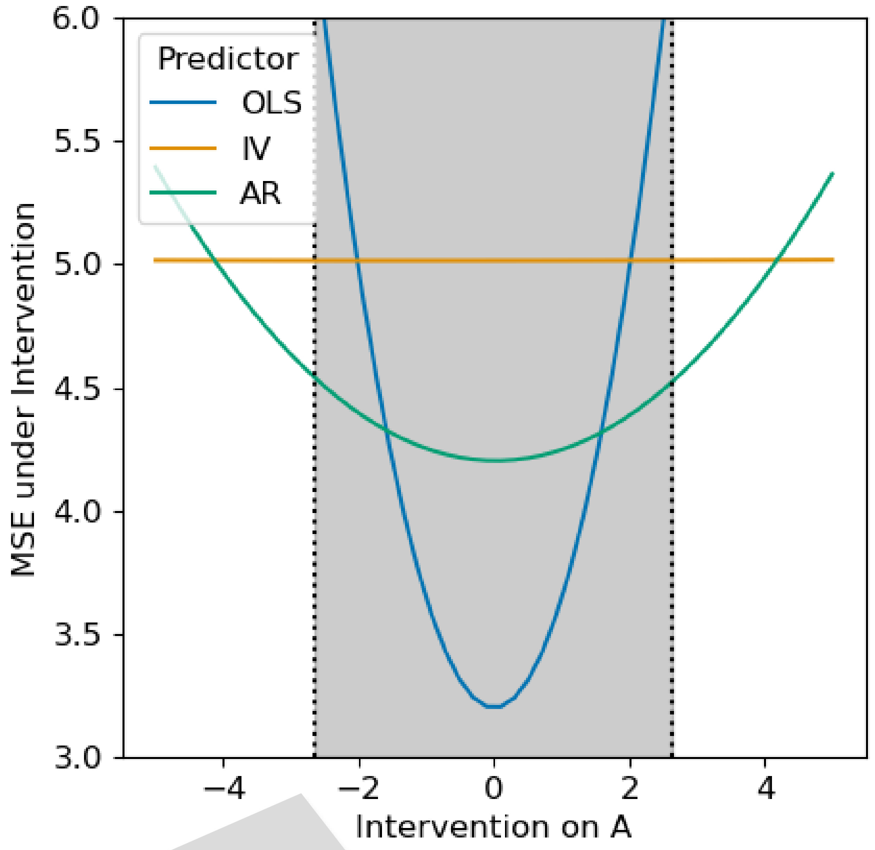
Optimize worst-case loss over interventions on A in a targeted robustness set.

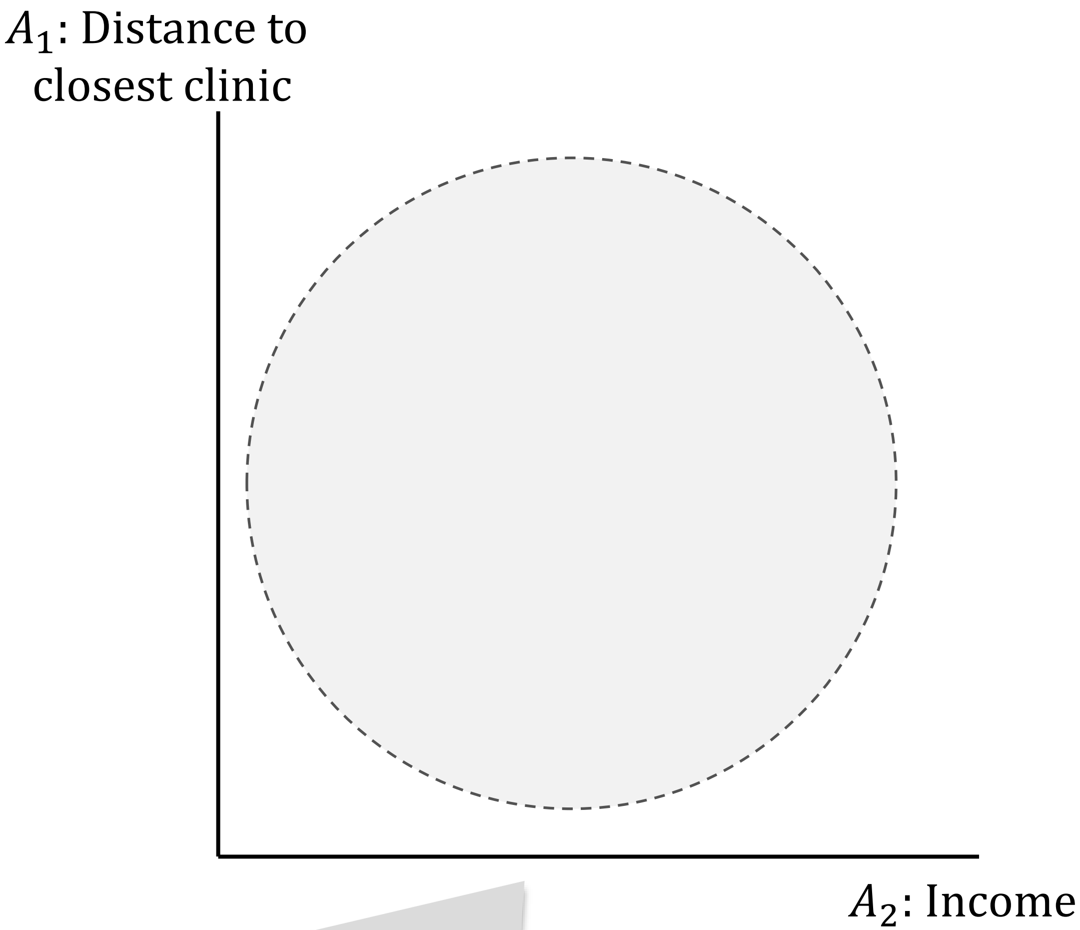
Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

2

Two proxies suffice to recover guarantees as if A were observed.

Aside: Considering multiple dimensions

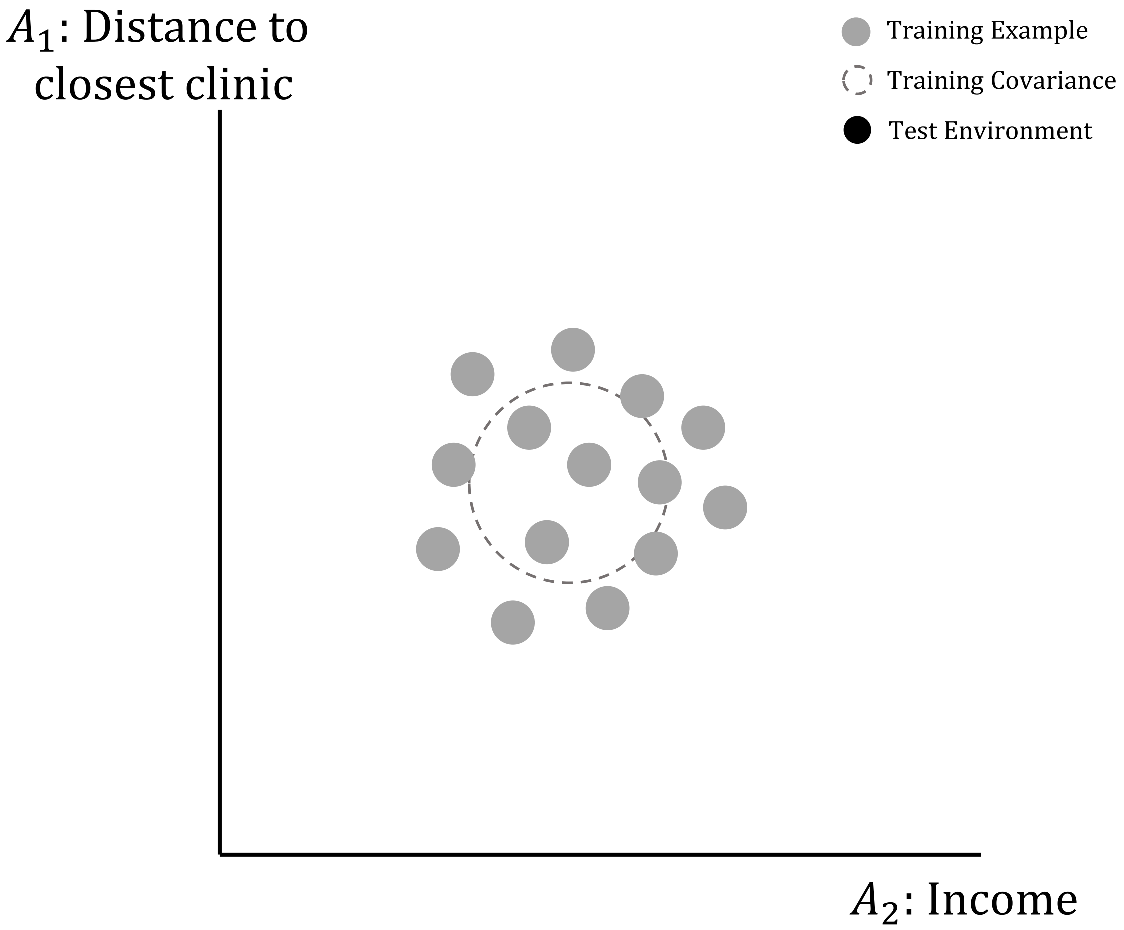




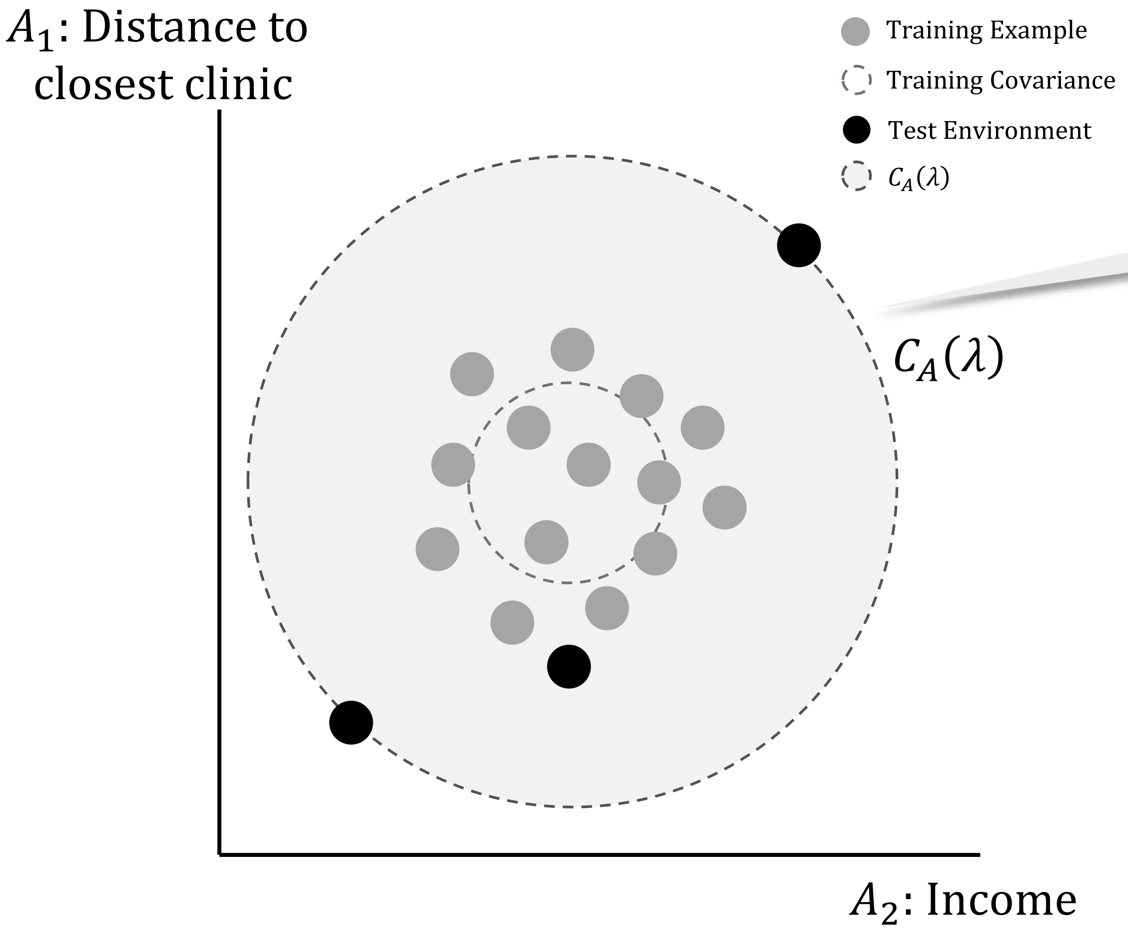
Robustness set in 1-dimension

Robustness set in multiple dimensions

Robustness to targeted interventions



Robustness to targeted interventions



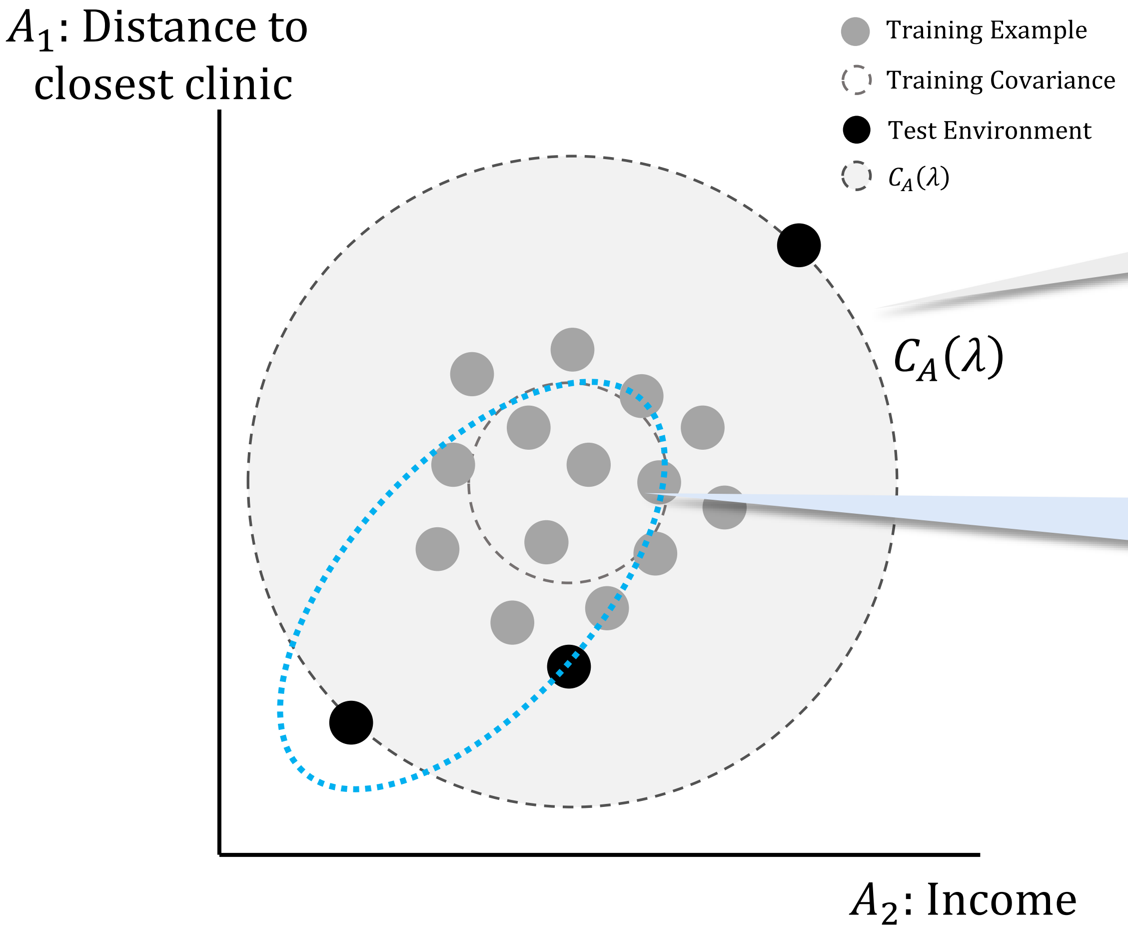
Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

$$C_A(\lambda) \coloneqq \{\nu : \nu\nu^\top \preccurlyeq (1+\lambda)\, \mathbb{E}[AA^\top]\}$$

[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.

Robustness to targeted interventions



Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

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

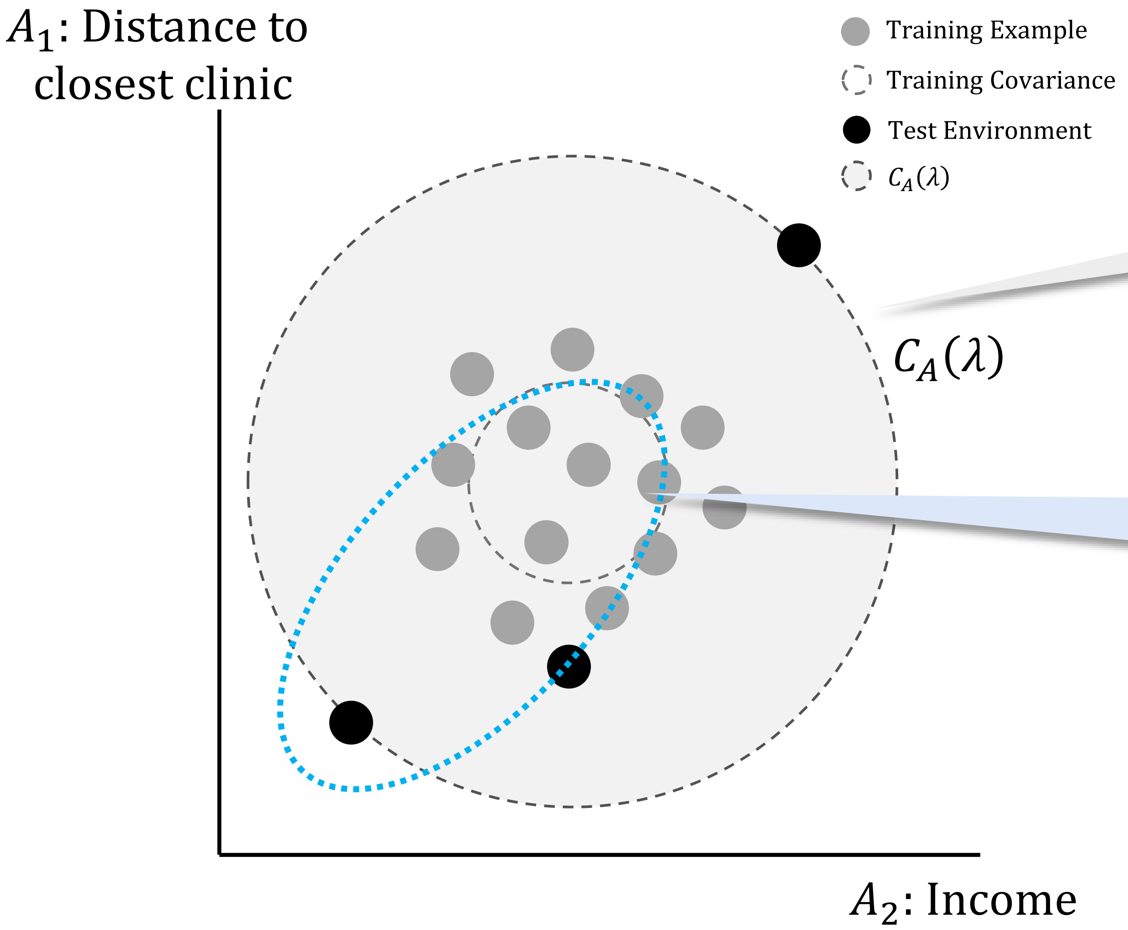
Targeted Anchor Regression optimizes a worst-case loss over an arbitrary ellipsoid robustness set

$$\sup_{\nu \in T(\mu_\nu, \Sigma_\nu)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X - \alpha)^2]$$

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : (\nu - \mu_\nu)(\nu - \mu_\nu)^\top \preccurlyeq \Sigma_\nu\}$$

[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.

Robustness to targeted interventions



Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

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

Targeted Anchor Regression optimizes a worst-case loss over an arbitrary ellipsoid robustness set

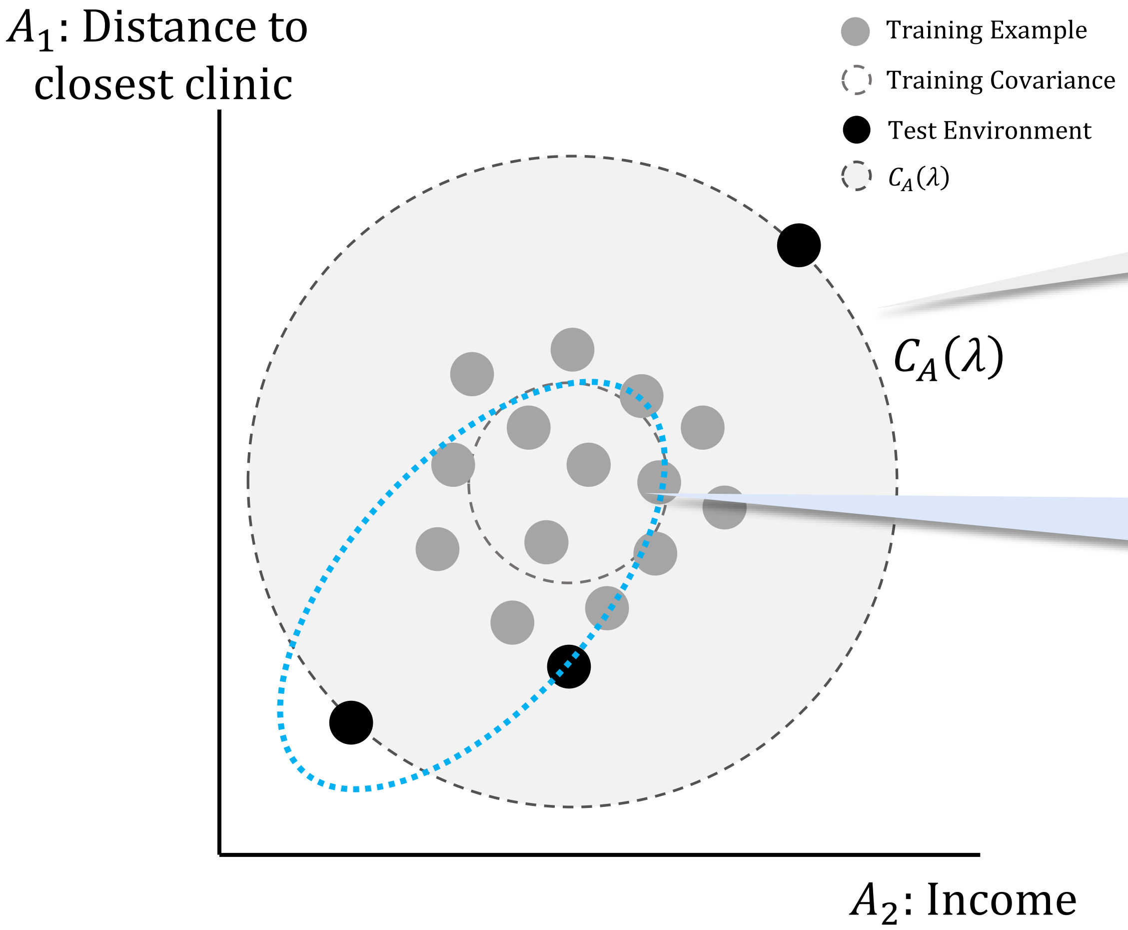
$$\sup_{\nu \in T(\mu_\nu, \Sigma_\nu)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X - \alpha)^2]$$

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : (\nu - \mu_\nu)(\nu - \mu_\nu)^\top \preccurlyeq \Sigma_\nu\}$$

Anchor Regression is a special case

[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.

Robustness to targeted interventions



Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

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

Targeted Anchor Regression optimizes a worst-case loss over an arbitrary ellipsoid robustness set

$$\sup_{\nu \in T(\mu_\nu, \Sigma_\nu)} \mathbb{E}_{do(A := \nu)}[(Y - \gamma^\top X - \alpha)^2]$$

$$T(\mu_\nu, \Sigma_\nu) := \{\nu : (\nu - \mu_\nu)(\nu - \mu_\nu)^\top \preccurlyeq \Sigma_\nu\}$$

Anchor Regression is a special case

Controls the **center** of the robustness set

Controls the **shape** of the robustness set

[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.

$$R(\gamma) := Y - \gamma^\top X$$

Targeted Anchor Regression

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \ell_{PLS}(X, Y, A; \gamma)$$

$$R(\gamma) := Y - \gamma^\top X$$

Targeted Anchor Regression

$$b_\gamma^\top := \mathbb{E}[R(\gamma)A^\top]\,(\mathbb{E}[AA^\top])^{-1}$$

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot b_\gamma^\top \mathbb{E}[AA^\top] b_\gamma$$

Intuition: Causal effect (on the residual) of a shift in A

$$R(\gamma) := Y - \gamma^\top X$$

Targeted Anchor Regression

Anchor Regression

$$b_\gamma^\top := \mathbb{E}[R(\gamma)A^\top]\,(\mathbb{E}[AA^\top])^{-1}$$

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot b_\gamma^\top \mathbb{E}[AA^\top] b_\gamma$$

Intuition: Causal effect (on the residual) of a shift in A

Targeted Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + b_\gamma^\top (\Sigma_\nu - \mathbb{E}[AA^\top]) b_\gamma + \left(b_\gamma^\top \mu_\nu - \alpha\right)^2$$

Controls the **shape** of the robustness set

Controls the **center** of the robustness set

Our contributions

1

Optimize worst-case loss over interventions on A in a **targeted** robustness set.

Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.

2

Two proxies suffice to recover guarantees as if A were observed.

Assumption: Noisy Proxies

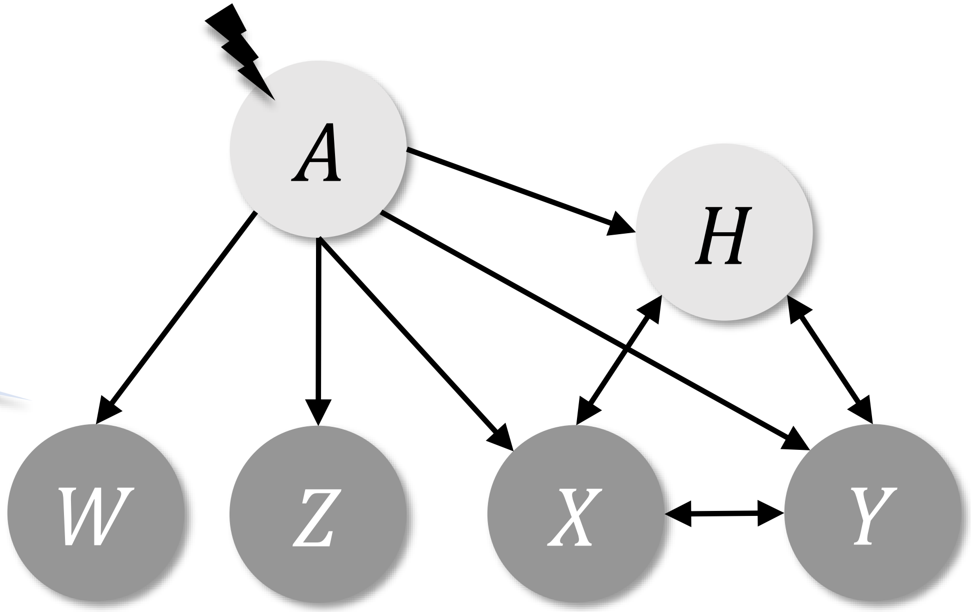
Assumptions: Linear structural causal model (SCM) over all observed and unobserved variables and **one or more noisy proxies of A**

Proxies are linear functions of A with independent additive noise.

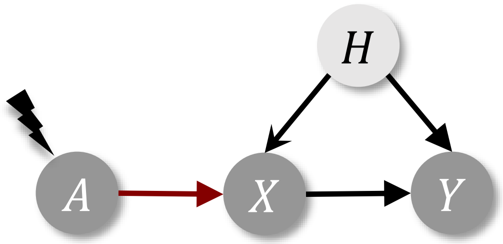
Example: Self-reported data on income, distance to closest clinic, etc.

$$W \coloneqq \beta_W A + \epsilon_W$$

$$Z \coloneqq \beta_Z A + \epsilon_Z$$

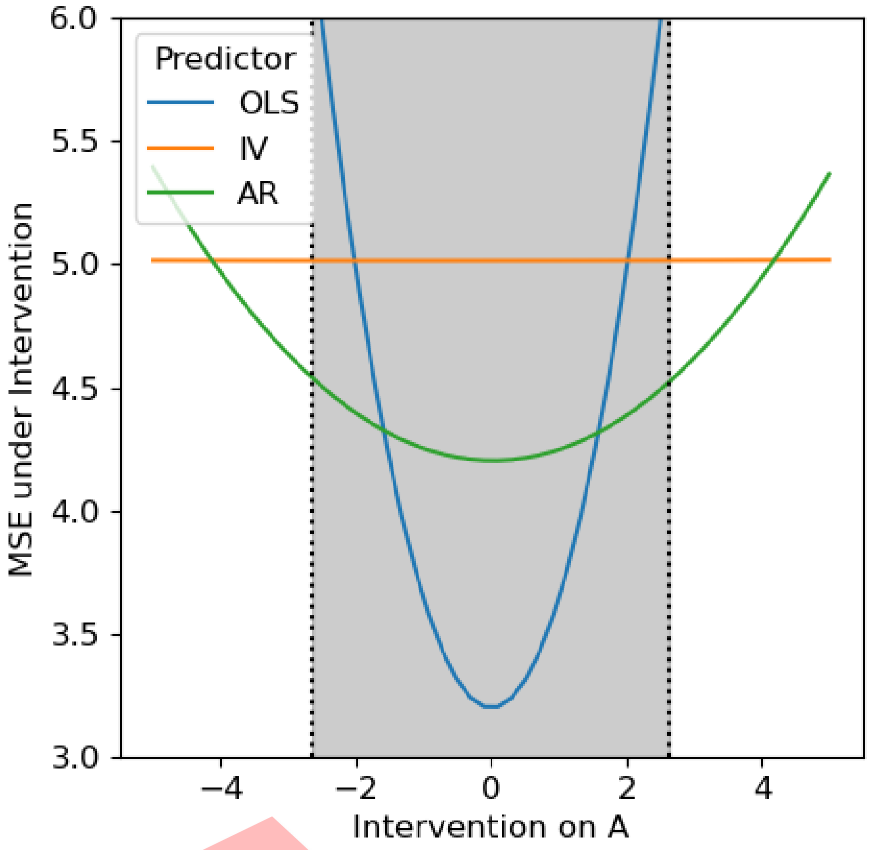


Single Proxy: Impact of Noise in 1D



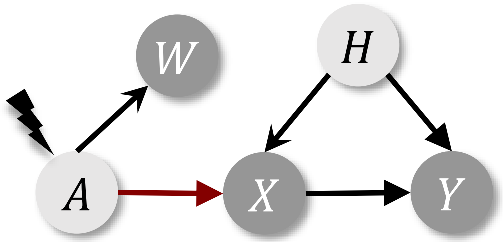
When **A is observed directly**, Anchor Regression minimizes the worst-case over

$$|\nu| < \sqrt{1 \ + \lambda}$$



Without noise, optimal worst-case risk for interventions on A up to $|\nu| < \sqrt{1+\lambda}$

Single Proxy: Impact of Noise in 1D

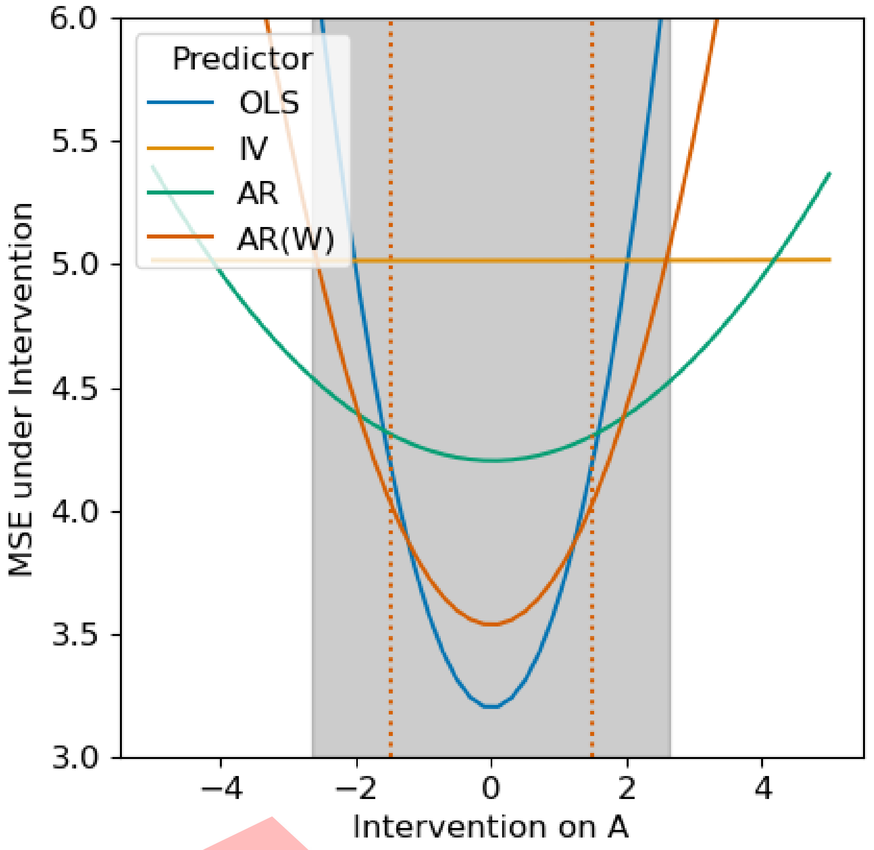


When **A is observed directly**, Anchor Regression minimizes the worst-case over

$$|\nu| < \sqrt{1 + \lambda}$$

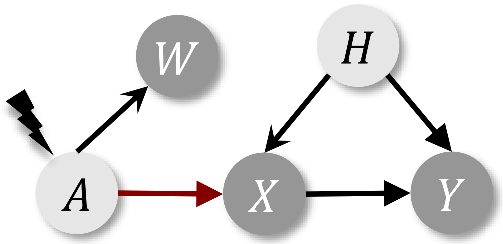
Using **a single noisy proxy W** in place of A, this robustness set becomes

$$|\nu| < \sqrt{1 + \lambda \cdot \rho_W}$$



With noisy W, optimal worst-case risk for interventions on A up to $|\nu| < \sqrt{1 + \boldsymbol{\lambda} \cdot \boldsymbol{\rho_W}}$

Single Proxy: Impact of Noise in 1D



When **A is observed directly**, Anchor Regression minimizes the worst-case over

$$|\nu| < \sqrt{1 + \lambda}$$

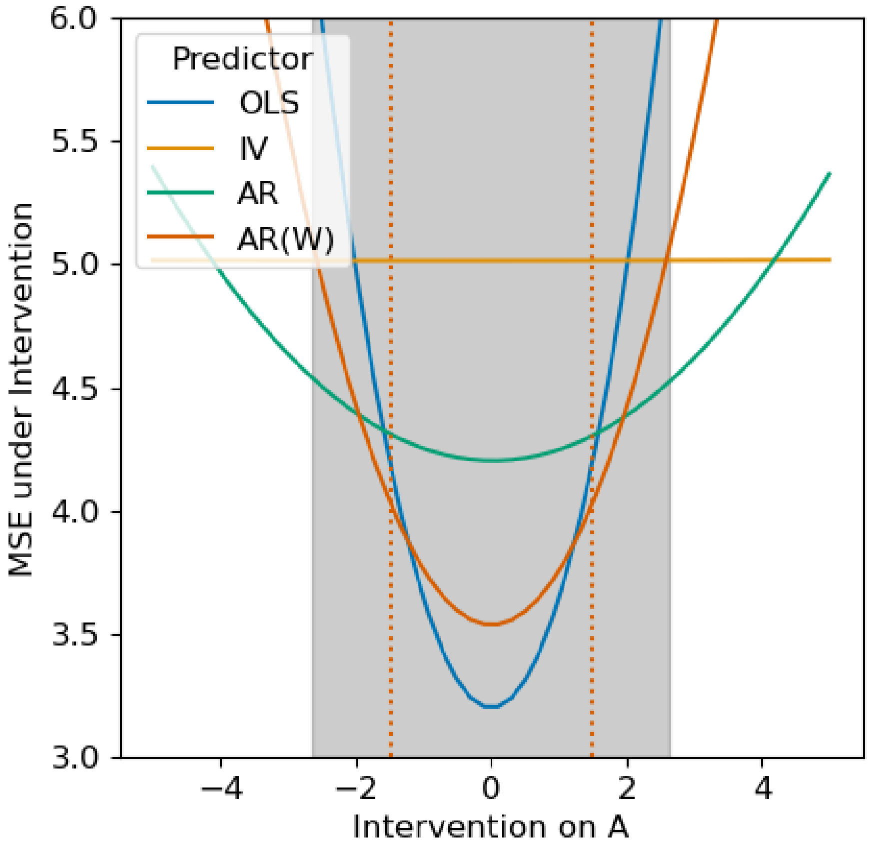
Using **a single noisy proxy W** in place of A, this robustness set becomes

$$|\nu| < \sqrt{1 + \lambda \cdot \rho_w}$$

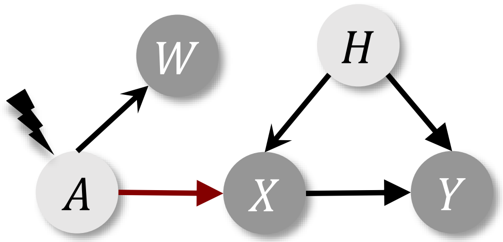
Where ρ_w is the **signal-to-variance ratio**

$$\rho_w = \frac{\beta_W^2}{\beta_W^2 + \mathbb{E}[\epsilon_w^2]} < 1$$

$$W := \beta_W A + \epsilon_W$$



Single Proxy: Impact of Noise in 1D



When **A is observed directly**, Anchor Regression minimizes the worst-case over

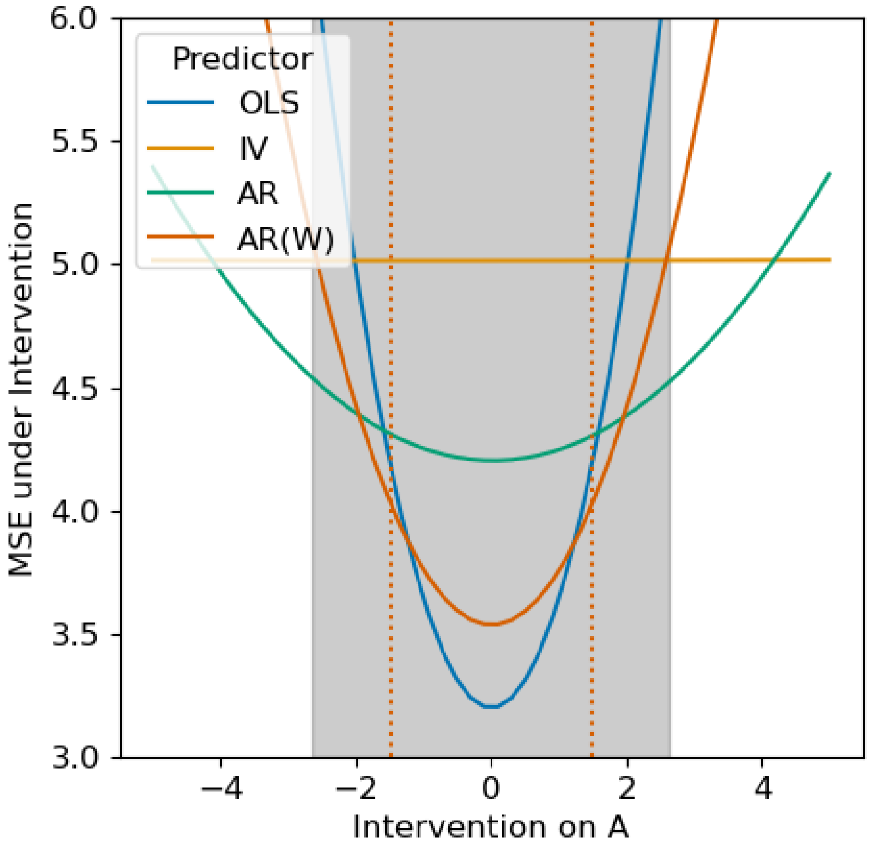
$$|\nu| < \sqrt{1 + \lambda}$$

Using **a single noisy proxy W** in place of A, this robustness set becomes

$$|\nu| < \sqrt{1 + \lambda \cdot \rho_w}$$

Where ρ_w is the **signal-to-variance ratio**

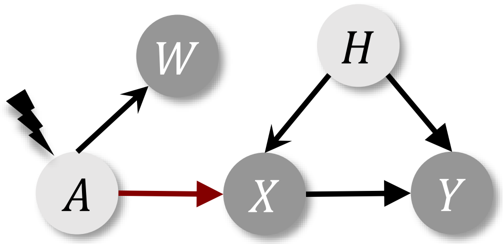
$$\rho_w = \frac{\beta_W^2}{\beta_W^2 + \mathbb{E}[\epsilon_w^2]} < 1$$



Value of ρ_W not generally identifiable from data!

$$W := \beta_W A + \epsilon_W$$

Single Proxy: Impact of Noise in 1D

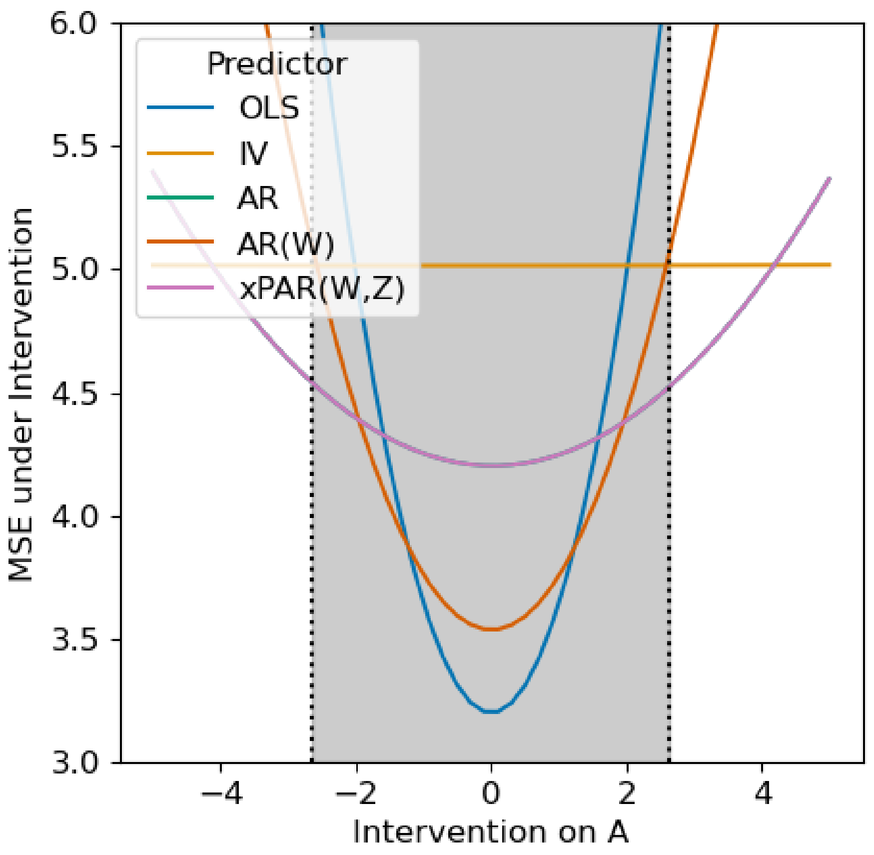


When **A is observed directly**, Anchor Regression minimizes the worst-case over

$$|\nu| < \sqrt{1 + \lambda}$$

Using **two noisy proxies** of A, we can recover the original guarantee

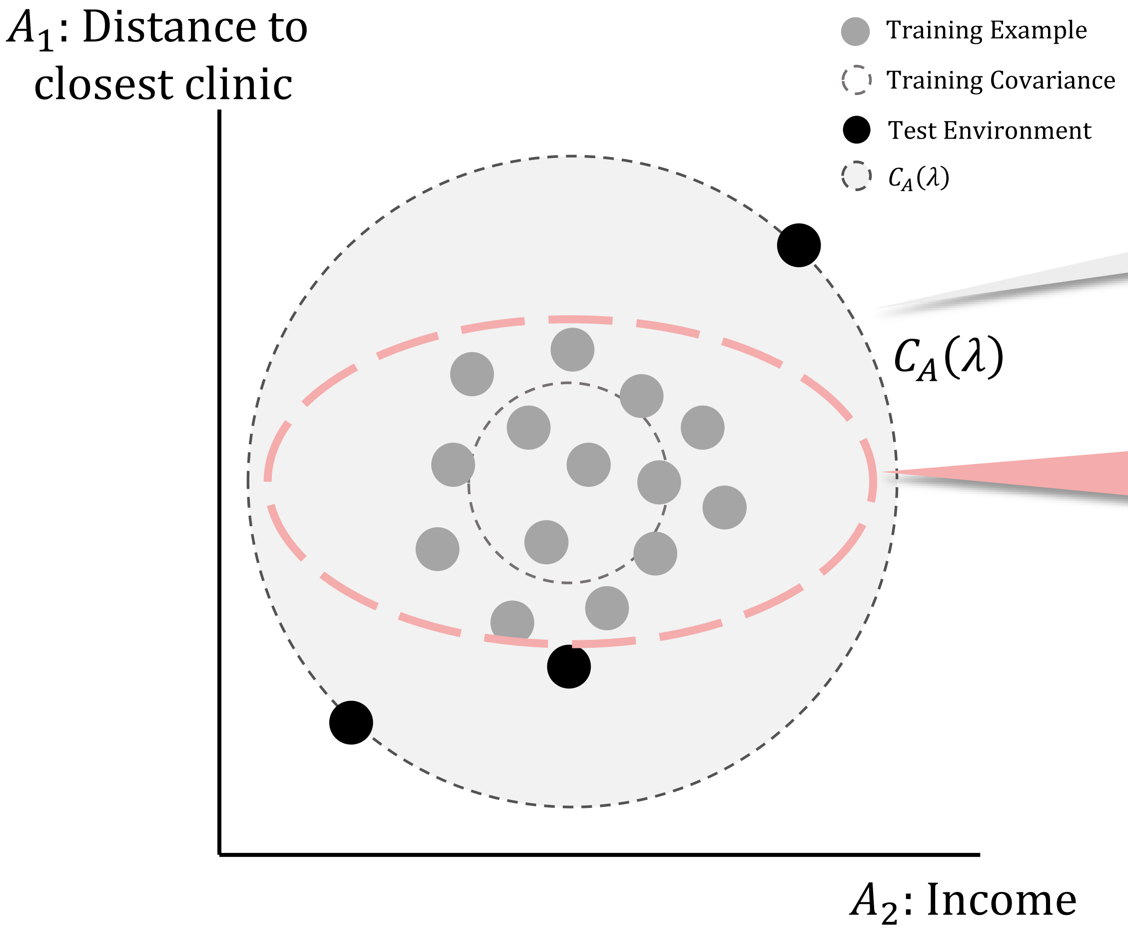
$$|\nu| < \sqrt{1 + \lambda}$$



$$Z := \beta_Z A + \epsilon_Z$$
$$W := \beta_W A + \epsilon_W$$

Requirement: β_Z, β_W are non-zero!

Impact of proxy noise in higher dimensions



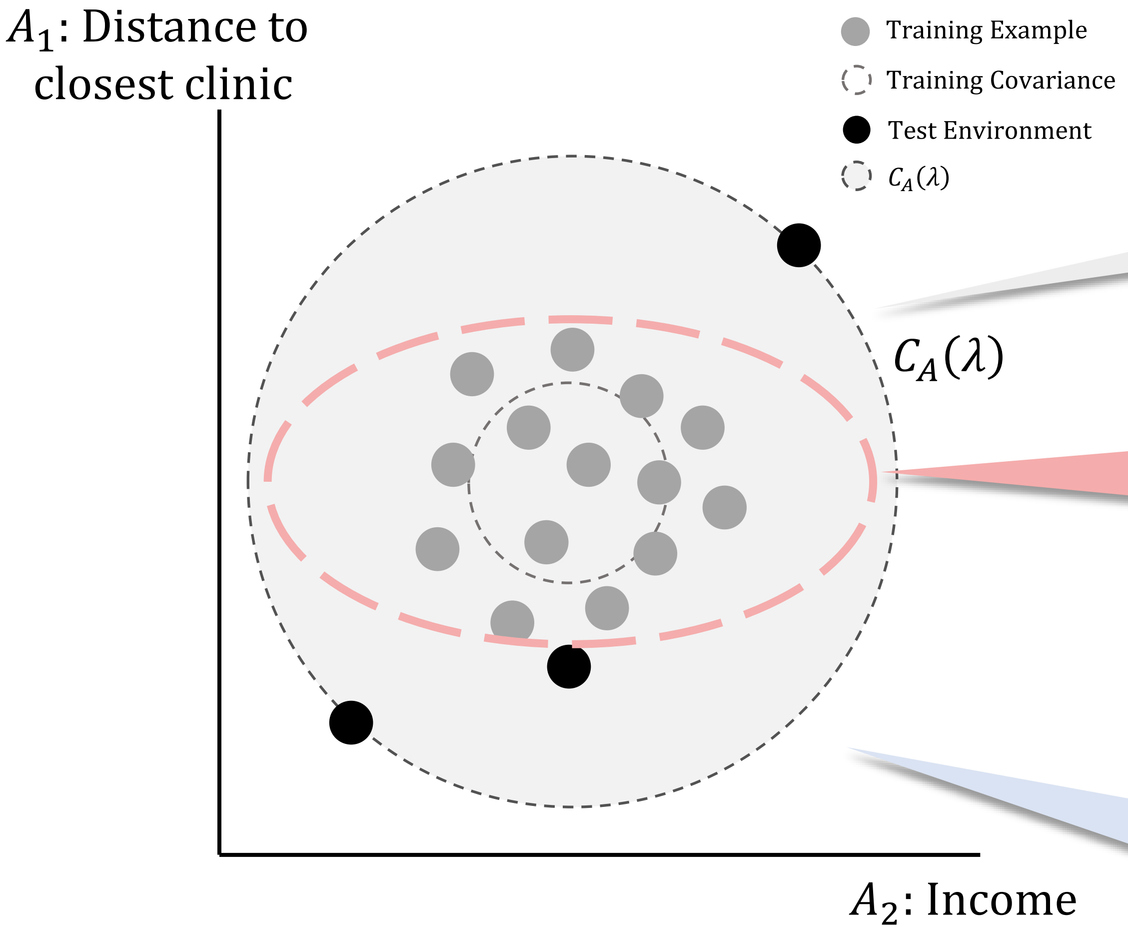
Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$$

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.

[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.

Impact of proxy noise in higher dimensions



Anchor Regression [1] optimizes a worst-case loss over interventions in a rescaling of the covariance of A

$$\sup_{\nu \in C_A(\lambda)} \mathbb{E}_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$$

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

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

[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.

$$R(\gamma) := Y - \gamma^\top X$$

Cross-Proxy Anchor Regression

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \ell_{PLS}(X, Y, A; \gamma)$$

$$R(\gamma) := Y - \gamma^\top X$$

Cross-Proxy Anchor Regression

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \ell_{PLS}(X, Y, A; \gamma)$$

Cross-Proxy Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \ell_{\times}(X, Y, W, Z; \gamma)$$

$$W := \beta_W A + \epsilon_W$$

$$Z := \beta_Z A + \epsilon_Z$$

$$R(\gamma) := Y - \gamma^\top X$$

Cross-Proxy Anchor Regression

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \mathbb{E}[R(\gamma)A^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AR(\gamma)^\top]$$

Cross-Proxy Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZR(\gamma)^\top]$$

$$W := \beta_W A + \epsilon_W$$

$$Z := \beta_Z A + \epsilon_Z$$

$$R(\gamma) := Y - \gamma^\top X$$

Cross-Proxy Anchor Regression

Anchor Regression

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \mathbb{E}[R(\gamma)A^\top]\mathbb{E}[AA^\top]^{-1}\mathbb{E}[AR(\gamma)^\top]$$

Cross-Proxy Anchor Regression

Equivalent to the fully-observed term, assuming β_W, β_Z full rank

$$\ell_{LS}(X, Y; \gamma) + \lambda \cdot \mathbb{E}[R(\gamma)W^\top]\mathbb{E}[ZW^\top]^{-1}\mathbb{E}[ZR(\gamma)^\top]$$

$$W := \beta_W A + \epsilon_W$$

$$Z := \beta_Z A + \epsilon_Z$$

We also extend this idea to Targeted AR, allowing for identification of more general worst-case

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
AR (TempC)	0.531			
TAR (TempC)	0.525			

Targeted AR uses knowledge of mean/variance of TempC in held-out

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536			
OLS + Est. Bias	0.569			
AR (TempC)	0.531			
TAR (TempC)	0.525			

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
AR (TempC)	0.531	6	-0.041	0.006
TAR (TempC)	0.525	8	-0.061	0.001

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
AR (TempC)	0.531	6	-0.041	0.006
PAR (W)				
xPAR (W, Z)				
TAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)				
xPTAR (W, Z)				

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
AR (TempC)	0.531	6	-0.041	0.006
PAR (W)				
xPAR (W, Z)				
TAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)	0.529	8	-0.038	0.001
xPTAR (W, Z)	0.526	7	-0.059	0.001

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Real Data Experiment: Pollution

Task

Predict pollution (PM2.5) based on weather-related variables. Data for five cities in China.

Setup

We use Temperature (Celsius) as our anchor variable, given seasonal effects.

Train/Validate: Train on 3 seasons, using leave-one-season-out CV to tune.

Evaluation: Predict on a held-out season, evaluate MSE.

Conclusion

1) Improves on OLS on average across evaluations, with limited downside.

2) Proxy noise hurts, but cross-proxy variants help mitigate the effect.

Table 1. Mean: Average MSE (lower is better) over 9 scenarios where $\lambda > 0$. # Win: Number of scenarios where the estimator has lower MSE than OLS. Best (Worst): Smallest (Largest) difference to OLS across environments, where lower is better.

Estimator	Mean	# Win	Best	Worst
OLS	0.537			
OLS (TempC)	0.536	5	-0.028	0.026
OLS + Est. Bias	0.569	4	-0.072	0.150
AR (TempC)	0.531	6	-0.041	0.006
PAR (W)	0.531	6	-0.037	0.006
xPAR (W, Z)	0.531	6	-0.039	0.007
TAR (TempC)	0.525	8	-0.061	0.001
PTAR (W)	0.529	8	-0.038	0.001
xPTAR (W, Z)	0.526	7	-0.059	0.001

[1] Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S. X. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. Journal of Geophysical Research: Atmospheres, 121, 2016.

Conclusions

Learn linear predictors that are ***robust to plausible interventions*** *on unobserved variables, using* ***noisy proxies*** *at training time.*

Conclusions

Learn linear predictors that are ***robust to plausible interventions*** *on unobserved variables, using* ***noisy proxies*** *at training time.*

Broader goal: Constructing domain-specific robustness guarantees

- Specifying relevant (unobserved) causal factors via proxies
- Focusing on plausible shifts in these underlying factors

Conclusions

*Learn linear predictors that are **robust to plausible interventions** on unobserved variables, using **noisy proxies** at training time.*

Broader goal: Constructing domain-specific robustness guarantees

- Specifying relevant (unobserved) causal factors via proxies
- Focusing on plausible shifts in these underlying factors

Open Directions: Extending to general (nonlinear) causal models.