Causal Effect Estimation with Context and Confounders (Pt. 2)

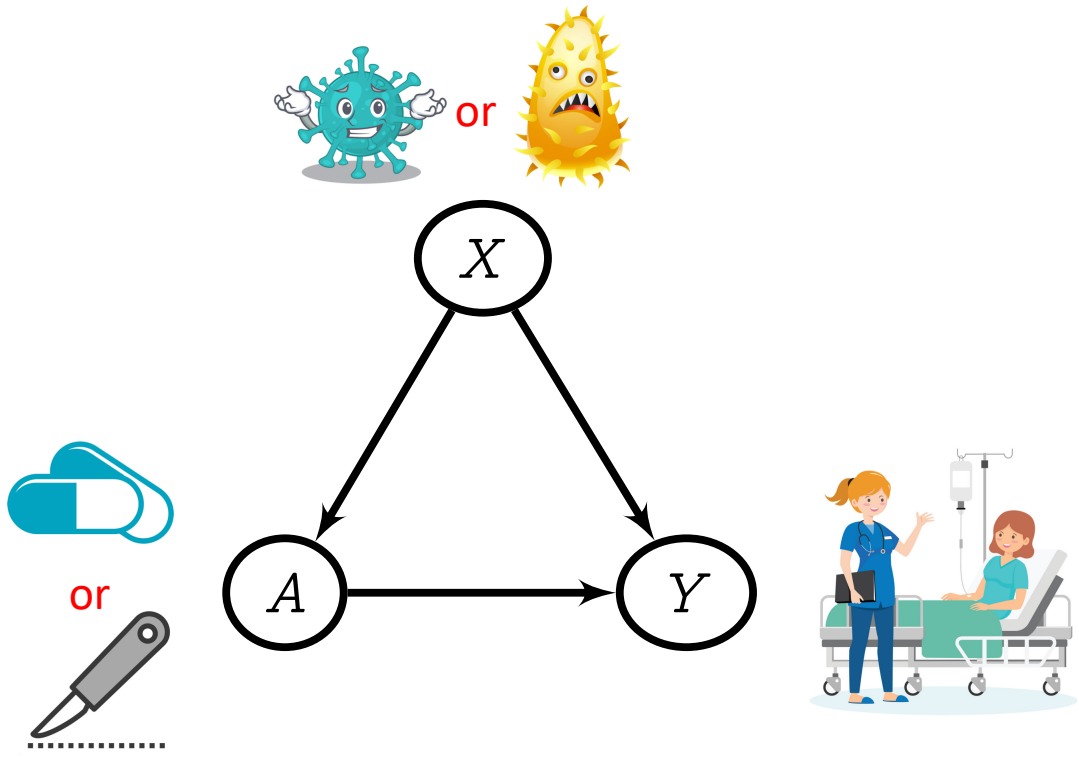
Arthur Gretton

Gatsby Computational Neuroscience Unit, UCL
Google DeepMind

MLSS, Tübingen 2026

Observation vs intervention

Conditioning from observation: $\mathbb{E}[Y|A=a] = \sum_x \mathbb{E}[Y|a,x]p(x|a)$

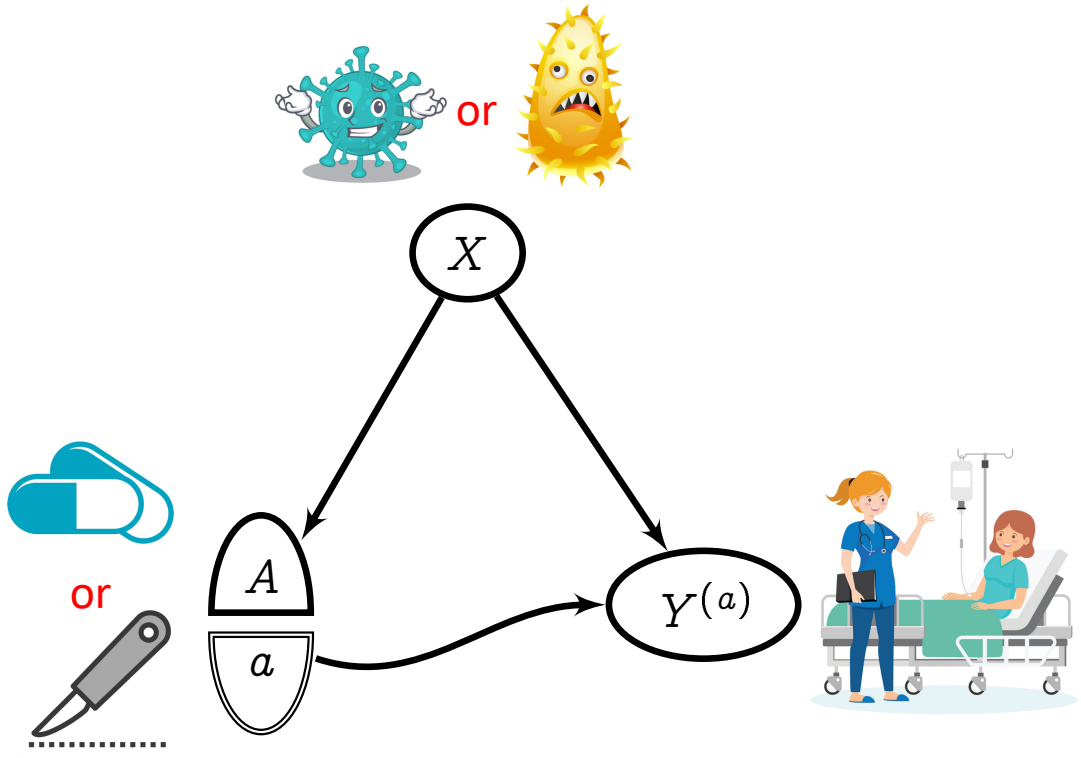


From our *observations* of historical hospital data:

- $P(Y = \text{cured}|A = \text{pills}) = 0.85$
- $P(Y = \text{cured}|A = \text{surgery}) = 0.72$

Observation vs intervention

Dose-reesponse curve (intervention): $\mathbb{E}[Y^{(a)}] = \sum_x \mathbb{E}[Y|a,x]p(x)$

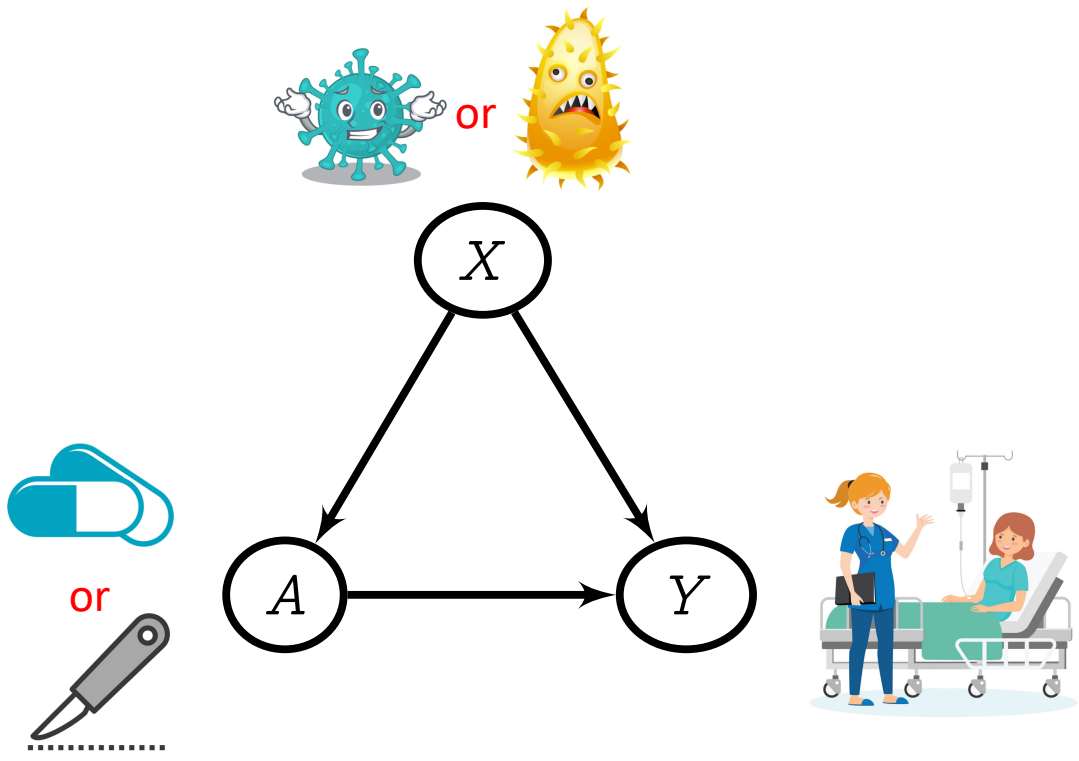


From our *intervention* (making all patients take a treatment):

- $P(Y^{(\text{pills})} = \text{cured}) = 0.64$
- $P(Y^{(\text{surgery})} = \text{cured}) = 0.75$

Richardson, Robins (2013), Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

Some core assumptions



Assume:

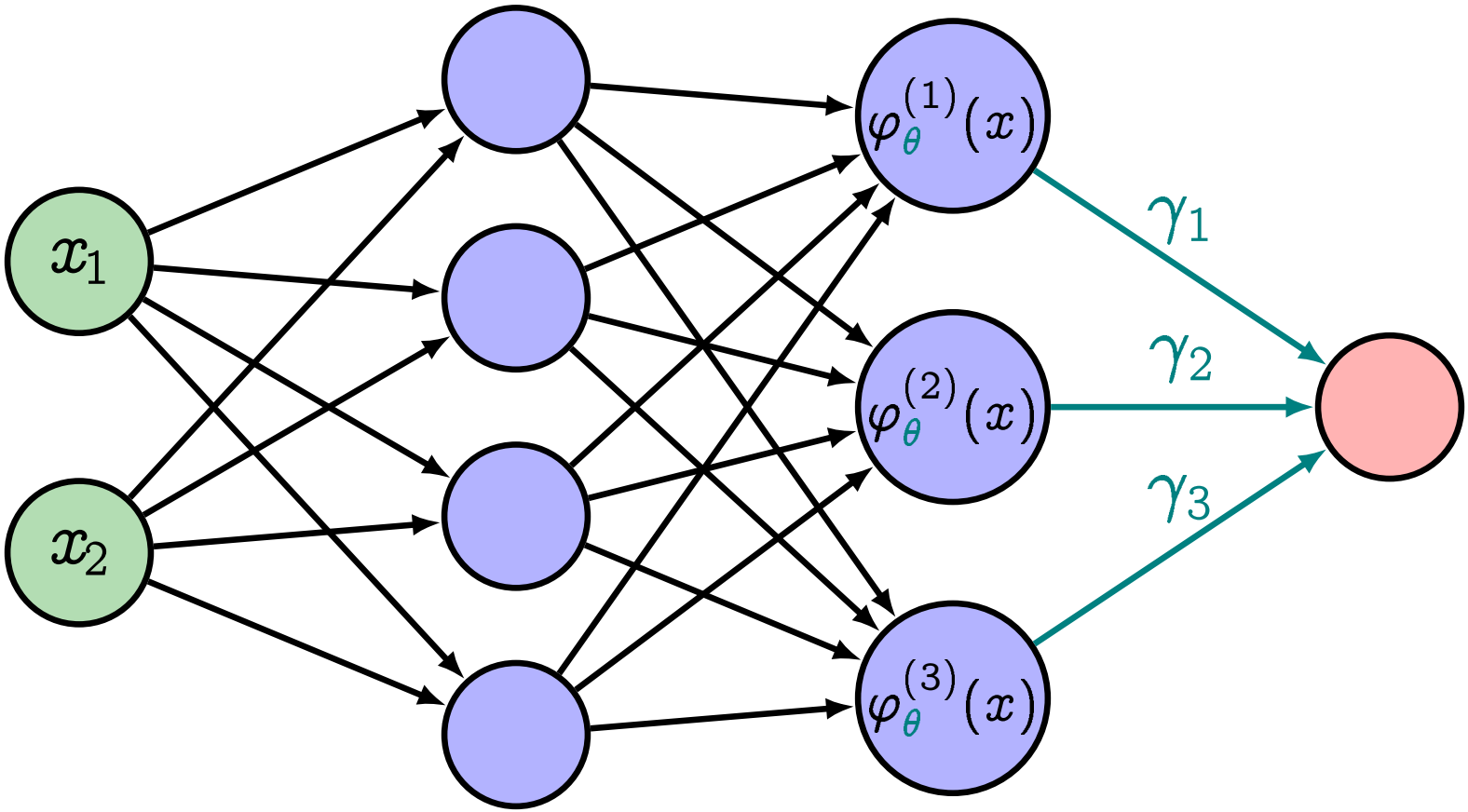
- Stable Unit Treatment Value Assumption (aka "no interference"),
- Conditional exchangeability $Y^{(a)} \perp\!\!\!\perp A|X$.
- Overlap.

One model: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x)$$

NN approach: Finite dictionaries of learned neural net features $\varphi_\theta(x)$ (linear final layer γ)



Model fitting: *neural* ridge regression

Learn $\gamma_0(x) := \mathbb{E}[Y|X = x]$ from features $\varphi_\theta(x_i)$ with outcomes y_i:

$$\hat{\gamma}_\theta \quad = \quad \arg\min_{\gamma} \left(\sum_{i=1}^{n} \left(y_i - \gamma^\top \varphi_\theta(x_i) \right)^2 + \lambda \|\gamma\|^2 \right) \tag{1}$$

Model fitting: *neural* ridge regression

Learn $\gamma_0(x) := \mathbb{E}[\,Y|X = x]$ from features $\varphi_\theta(x_i)$ with outcomes y_i:

$$\hat{\gamma}_\theta \quad = \quad \arg\min_{\gamma} \left(\sum_{i=1}^{n} \left(y_i - \gamma^\top \varphi_\theta(x_i) \right)^2 + \lambda \|\gamma\|^2 \right) \tag{1}$$

Solution for linear final layer γ:

$$\hat{\gamma}_\theta = C_{YX}^{(\theta)} (\, C_{XX}^{(\theta)} + \lambda)^{-1}$$

$$C_{YX}^{(\theta)} = \frac{1}{n} \sum_{i=1}^{n} [y_i \, \varphi_\theta(x_i)^\top]$$

$$C_{XX}^{(\theta)} = \frac{1}{n} \sum_{i=1}^{n} [\varphi_\theta(x_i) \, \varphi_\theta(x_i)^\top]$$

Model fitting: *neural* ridge regression

Learn $\gamma_0(x) := \mathbb{E}[\,Y|X = x]$ from features $\varphi_\theta(x_i)$ with outcomes y_i:

$$\hat{\gamma}_\theta \quad = \quad \arg\min_{\gamma} \left(\sum_{i=1}^{n} \left(y_i - \gamma^\top \varphi_\theta(x_i) \right)^2 + \lambda \|\gamma\|^2 \right) \tag{1}$$

Solution for linear final layer γ:

$$\hat{\gamma}_\theta = C_{YX}^{(\theta)} (C_{XX}^{(\theta)} + \lambda)^{-1}$$

$$C_{YX}^{(\theta)} = \frac{1}{n} \sum_{i=1}^{n} [y_i \, \varphi_\theta(x_i)^\top]$$

$$C_{XX}^{(\theta)} = \frac{1}{n} \sum_{i=1}^{n} [\varphi_\theta(x_i) \, \varphi_\theta(x_i)^\top]$$

How to solve for θ:

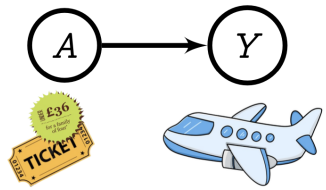
Substitute $\hat{\gamma}_\theta$ into (1), backprop for remaining θ.

More details: Galashov, Da Costa, Xu, Hennig, G, Closed-Form Last Layer Optimization (2025, arxiv:2510.04606)

Instrumental variable regression

Illustration: ticket prices for air travel

Ticket price A, seats sold Y.

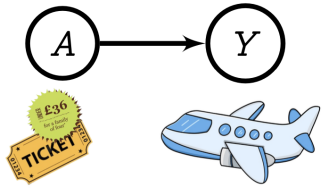


What is the effect on seats sold $Y^{(a)}$ of intervening on price a?

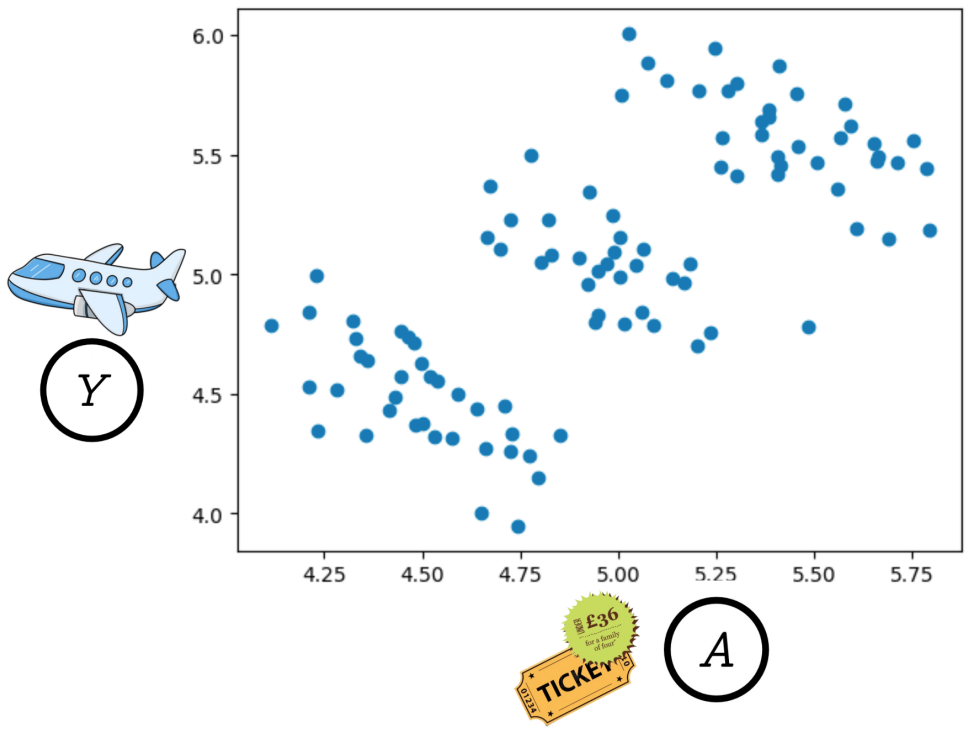
Simplification of example from Hartford, Lewis, Leyton-Brown, Taddy (2017): Deep IV: A Flexible Approach for Counterfactual Prediction.

Illustration: ticket prices for air travel

Ticket price A, seats sold Y.



What is the effect on seats sold $Y^{(a)}$ of intervening on price a?

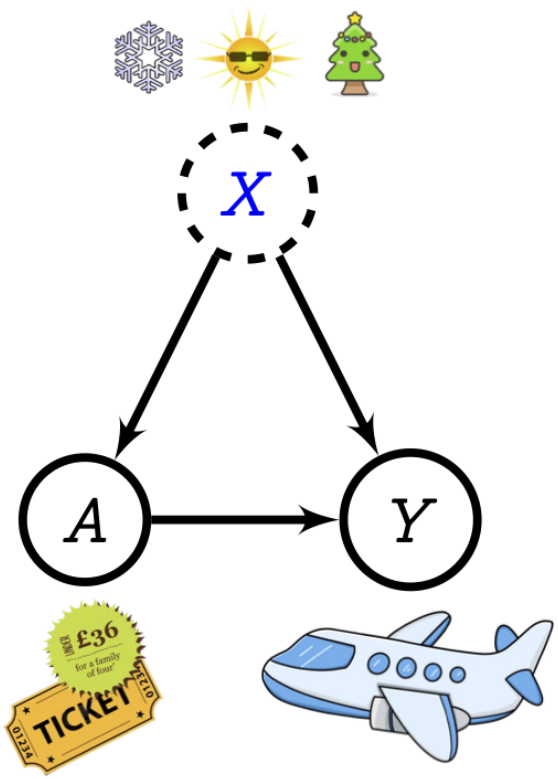


Simplification of example from Hartford, Lewis, Leyton-Brown, Taddy (2017): Deep IV: A Flexible Approach for Counterfactual Prediction.

Illustration: ticket prices for air travel

Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.

- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$



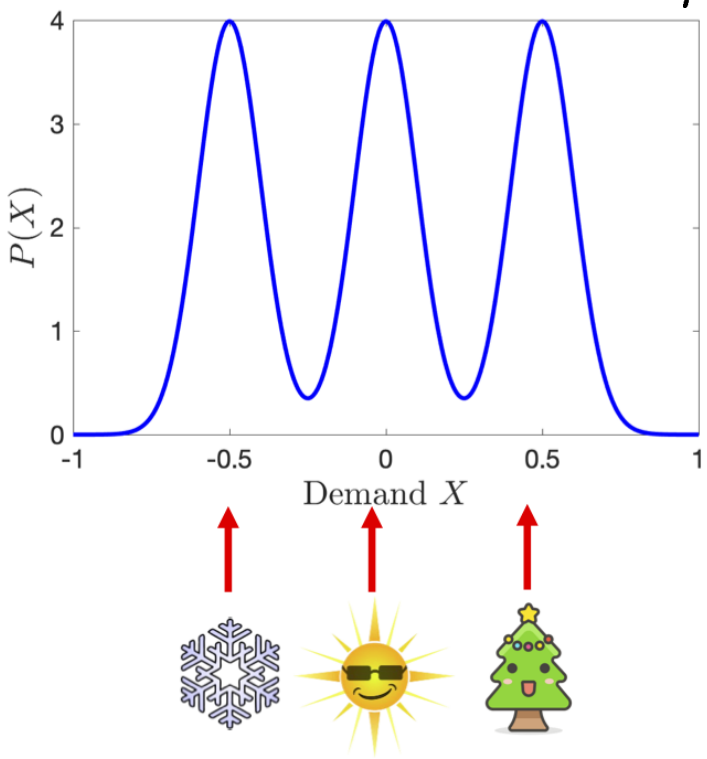
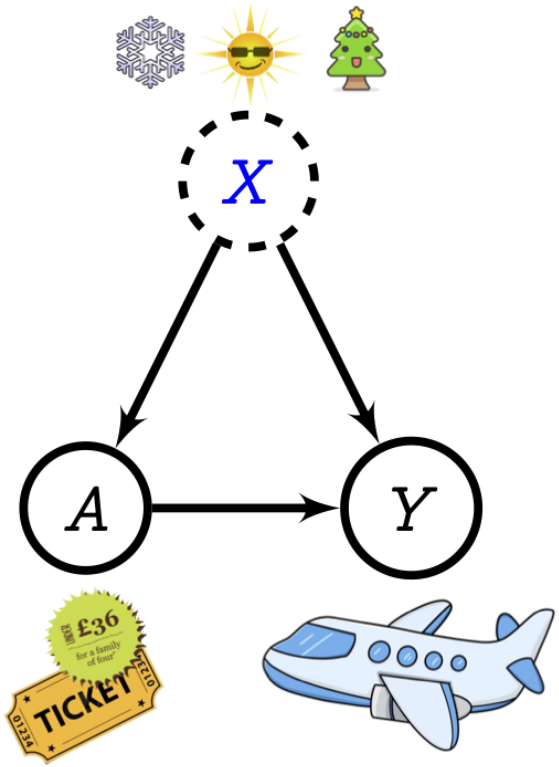


Illustration: ticket prices for air travel

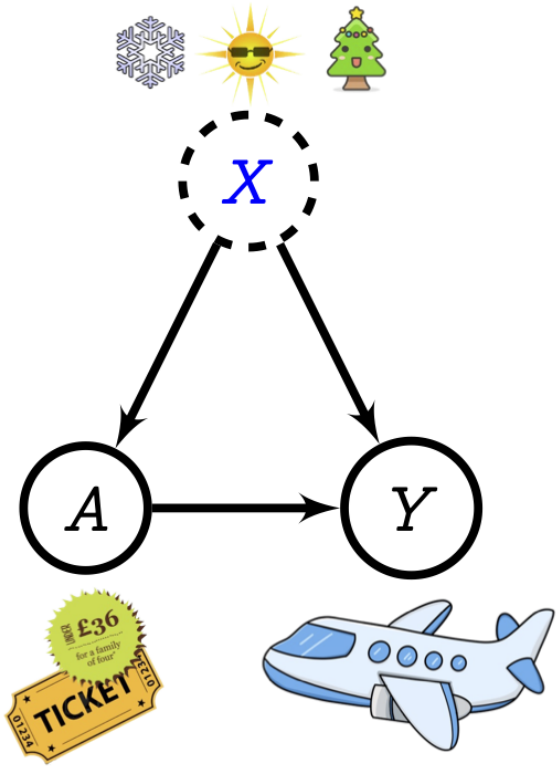
Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.



- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$
- Price:
  $A = X + Z,$
  $Z \sim \mathcal{N}(5, 0.04)$

Illustration: ticket prices for air travel

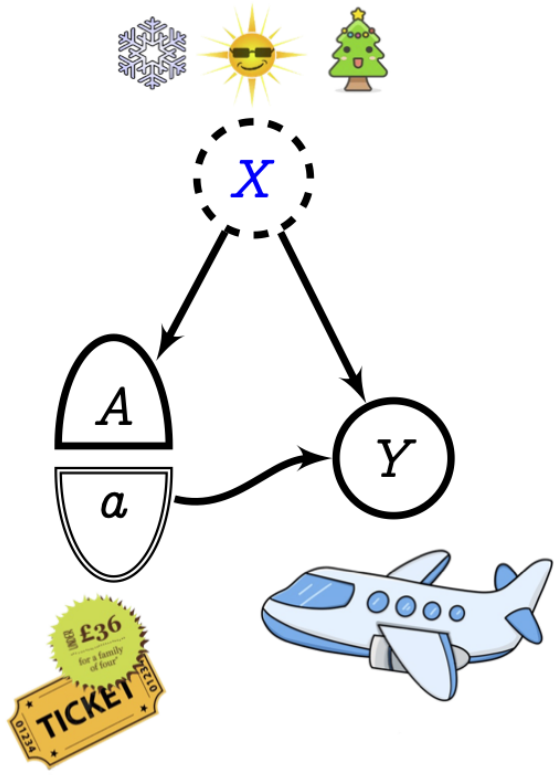
Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.



- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$
- Price:
  $A = X + Z,$
  $Z \sim \mathcal{N}(5, 0.04)$
- Seats sold:
  $Y = 10 - A + 2X$

Illustration: ticket prices for air travel

Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.



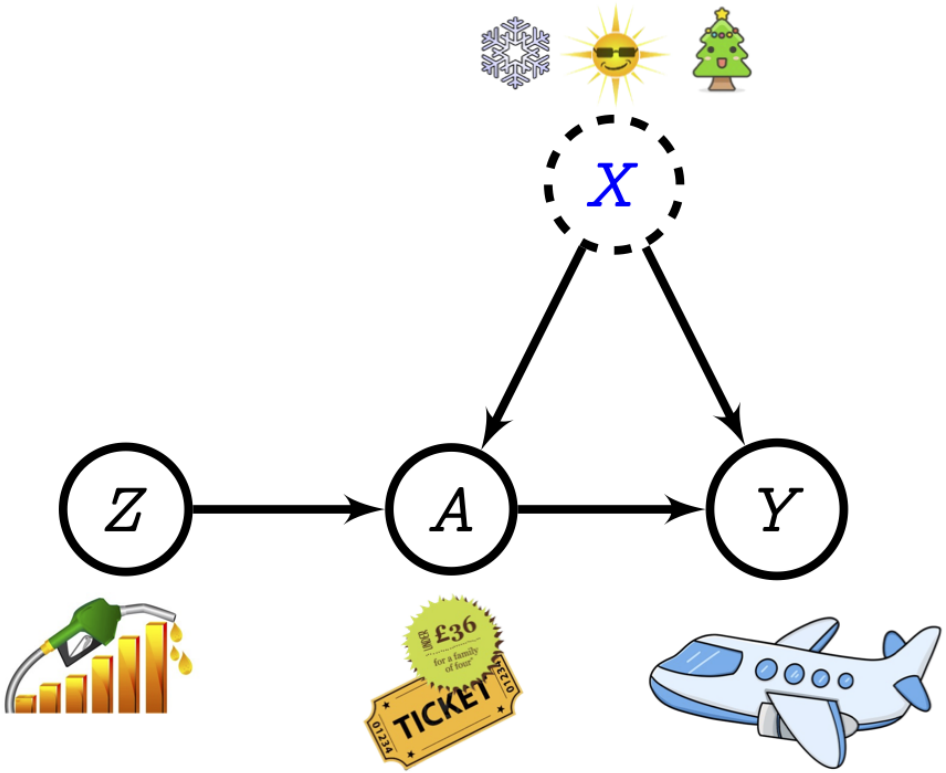
- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$
- Price:
  $A = X + Z,$
  $Z \sim \mathcal{N}(5, 0.04)$
- Seats sold:
  $Y = 10 - A + 2X$

Dose-response curve:

$$\mathrm{DR}(a) = \mathbb{E}[\,Y^{(a)}] = \int (10 - a + 2X)\, dp(X) = 10 - a$$

Illustration: ticket prices for air travel

Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.



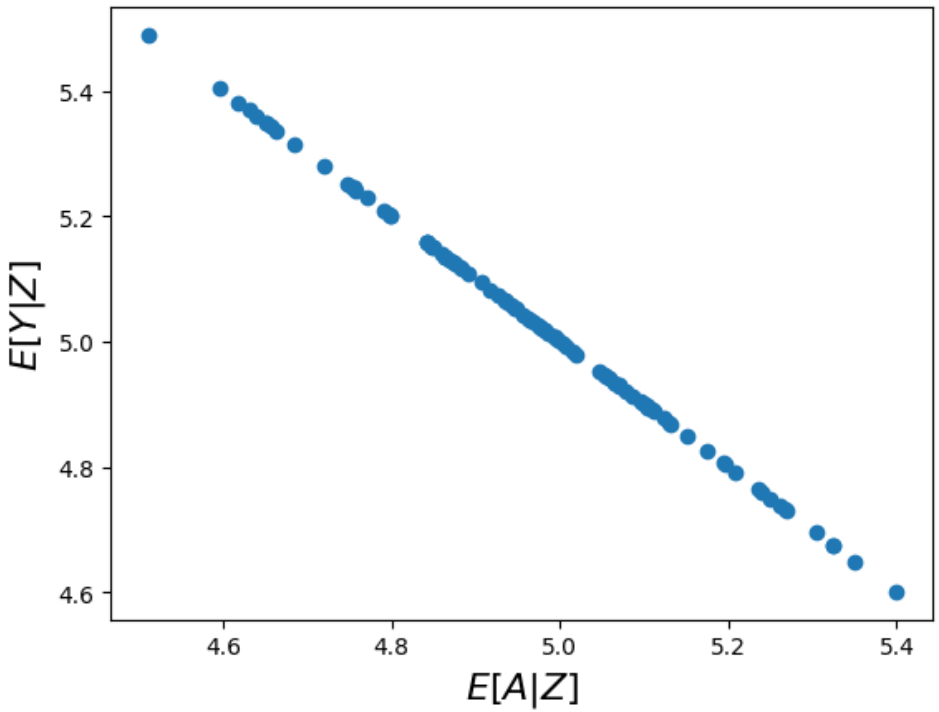
- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$
- Price:
  $A = X + Z,$
  $Z \sim \mathcal{N}(5, 0.04)$
- Seats sold:
  $Y = 10 - A + 2X$

Z is an instrument (cost of fuel). Condition on Z,

$$\mathbb{E}[Y|Z] = 10 - \mathbb{E}[A|Z] + 2\underbrace{\mathbb{E}[X|Z]}_{=0}$$

Illustration: ticket prices for air travel

Unobserved variable X =desire for travel, affects *both* price (via airline algorithms) *and* seats sold.



- Desire for travel:
  $X \sim \mathcal{N}(\mu, 0.01)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$
- Price:
  $A = X + Z,$
  $Z \sim \mathcal{N}(5, 0.04)$
- Seats sold:
  $Y = 10 - A + 2X$

Z is an instrument (cost of fuel). Condition on Z,

$$\mathbb{E}[Y|Z] = 10 - \mathbb{E}[A|Z] + 2\underbrace{\mathbb{E}[X|Z]}_{=0}$$

Regressing from $\mathbb{E}[A|Z]$ to $\mathbb{E}[Y|Z]$ recovers causal relation!

Instrumental variable regression

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

© Nobel Prize Outreach. Photo: Paul Kennedy

David Card

Prize share: 1/2

© Nobel Prize Outreach. Photo: Risdon Photography

Joshua D. Angrist

Prize share: 1/4

© Nobel Prize Outreach. Photo: Paul Kennedy

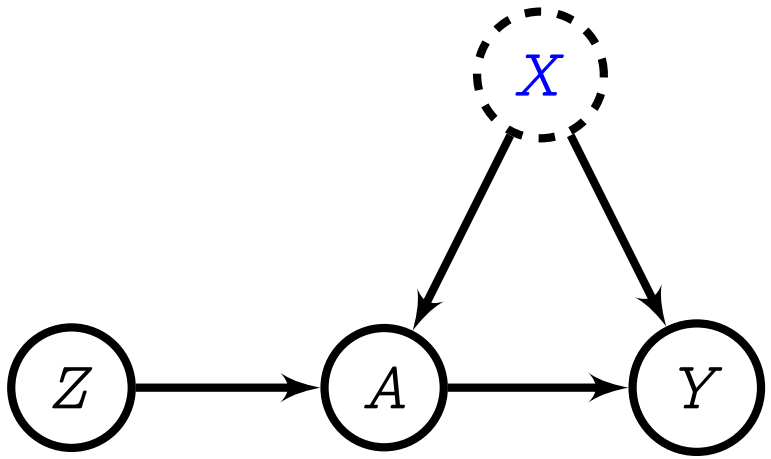
Guido W. Imbens

Prize share: 1/4

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 was divided, one half awarded to David Card "for his empirical contributions to labour economics", the other half jointly to Joshua D. Angrist and Guido W. Imbens "for their methodological contributions to the analysis of causal relationships"

Instrumental variable regression with NN features

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument



Assumptions

$$\mathbb{E}[X|Z] = 0$$

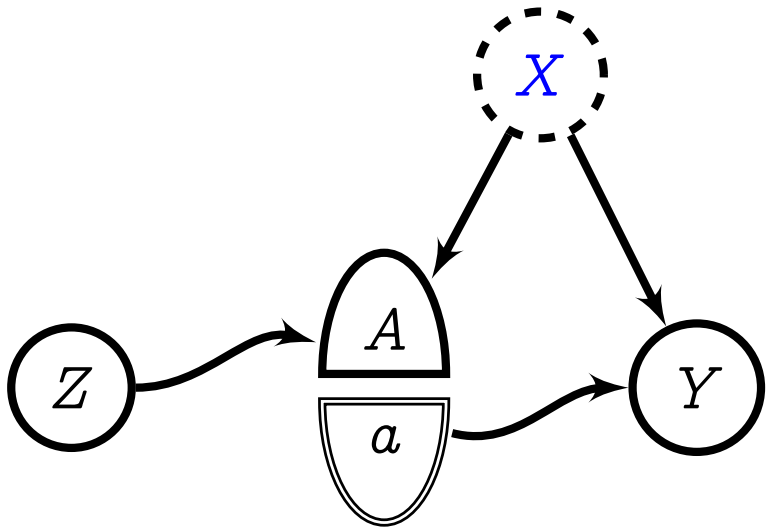
$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

$$Y = \gamma^\top \phi_\theta(A) + X$$

Instrumental variable regression with NN features

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument



Assumptions

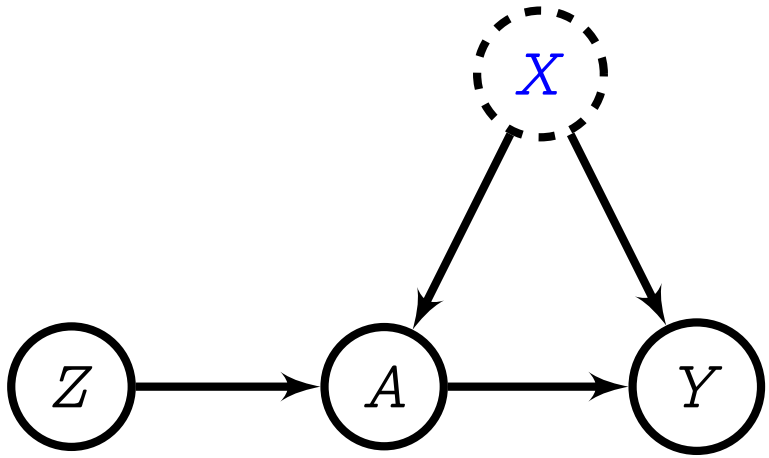
$$\mathbb{E}[X|Z] = 0$$
$$Z \not\perp\!\!\!\perp A$$
$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$
$$Y = \gamma^\top \phi_\theta(A) + X$$

Dose-response curve:

$$\mathrm{DR}(a) = \int \mathbb{E}(Y|X,a)\,dp(X) = \gamma^\top \phi_\theta(a)$$

Instrumental variable regression with NN features

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: instrument



Assumptions

$$\mathbb{E}[X|Z] = 0$$

$$Z \not\perp\!\!\!\perp A$$

$$(Y \perp\!\!\!\perp Z|A)_{G_{\bar{A}}}$$

$$Y = \gamma^\top \phi_\theta(A) + X$$

IV regression: Condition both sides on Z,

$$\mathbb{E}[Y|Z] = \gamma^\top \mathbb{E}[\phi_\theta(A)|Z] + \underbrace{\mathbb{E}[X|Z]}_{=0}$$

Plain linear regression: what goes wrong?

Output $y \in \mathbb{R}$, noise $X \in \mathbb{R}$, input A with NN features $\varphi_\theta(a)$.

Crucially, $X \not\perp\!\!\!\perp A$ and

$$C_{ax} := \mathbb{E}[\varphi_\theta(A)X] \neq 0$$

Plain linear regression: what goes wrong?

Output $y \in \mathbb{R}$, noise $X \in \mathbb{R}$, input A with NN features $\varphi_\theta(a)$.

Crucially, $X \not\perp\!\!\!\perp A$ and

$$C_{ax} := \mathbb{E}[\varphi_\theta(A)X] \neq 0$$

Dose-response curve:

$$y = {\gamma_0}^\top \varphi_\theta(a) + X \qquad \mathbb{E}(X) = 0$$

$$\mathrm{DR} := \mathbb{E}(Y^{(a)}) = \int ({\gamma_0}^\top \varphi_\theta(a) + X)\, dP(X) = {\gamma_0}^\top \varphi_\theta(a).$$

Plain linear regression: what goes wrong?

Output $y \in \mathbb{R}$, noise $X \in \mathbb{R}$, input A with NN features $\varphi_\theta(a)$.

Crucially, $X \not\perp\!\!\!\perp A$ and

$$C_{ax} := \mathbb{E}[\varphi_\theta(A)X] \neq 0$$

Dose-response curve:

$$y = {\gamma_0}^\top \varphi_\theta(a) + X \qquad \mathbb{E}(X) = 0$$

$$\mathrm{DR} := \mathbb{E}(Y^{(a)}) = \int ({\gamma_0}^\top \varphi_\theta(a) + X)\, dP(X) = {\gamma_0}^\top \varphi_\theta(a).$$

Problem: Least-squares loss for γ:

$$\mathcal{L}(\gamma, \theta) = \mathbb{E} \left\| Y - \gamma^\top \varphi_\theta(A) - X \right\|^2$$

Plain linear regression: what goes wrong?

Output $y \in \mathbb{R}$, noise $X \in \mathbb{R}$, input A with NN features $\varphi_\theta(a)$.

Crucially, $X \not\perp\!\!\!\perp A$ and

$$C_{ax} := \mathbb{E}[\varphi_\theta(A)X] \neq 0$$

Dose-response curve:

$$y = {\gamma_0}^\top \varphi_\theta(a) + X \qquad \mathbb{E}(X) = 0$$

$$\mathrm{DR} := \mathbb{E}(Y^{(a)}) = \int ({\gamma_0}^\top \varphi_\theta(a) + X)\, dP(X) = {\gamma_0}^\top \varphi_\theta(a).$$

Problem: Least-squares loss for γ:

$$\mathcal{L}(\gamma, \theta) = \mathbb{E} \left\| Y - \gamma^\top \varphi_\theta(A) - X \right\|^2$$

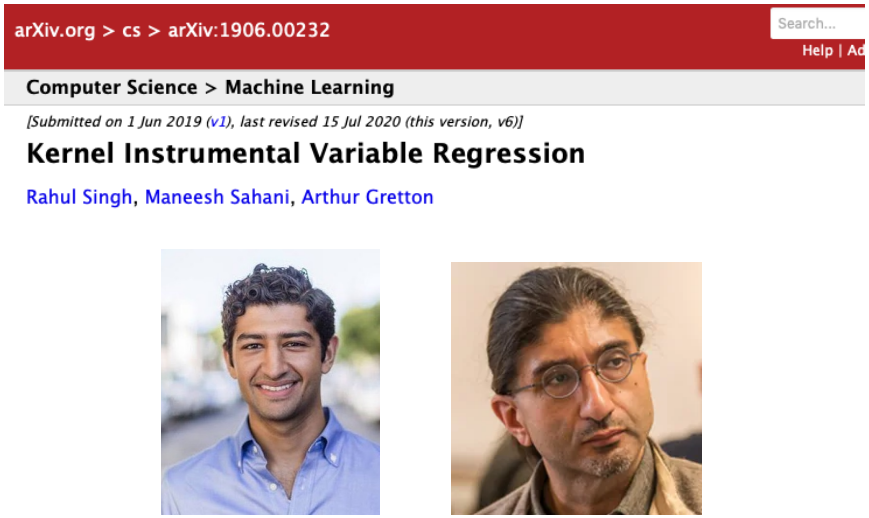
Minimizing for γ,

$$\gamma_0 = C_{aa}^{-1}(C_{ay} - C_{ax}) \qquad C_{aa} = \mathbb{E}[\varphi_\theta(A)\varphi_\theta(A)^\top]$$
$$C_{ay} = \mathbb{E}[\varphi_\theta(A)\, Y]$$

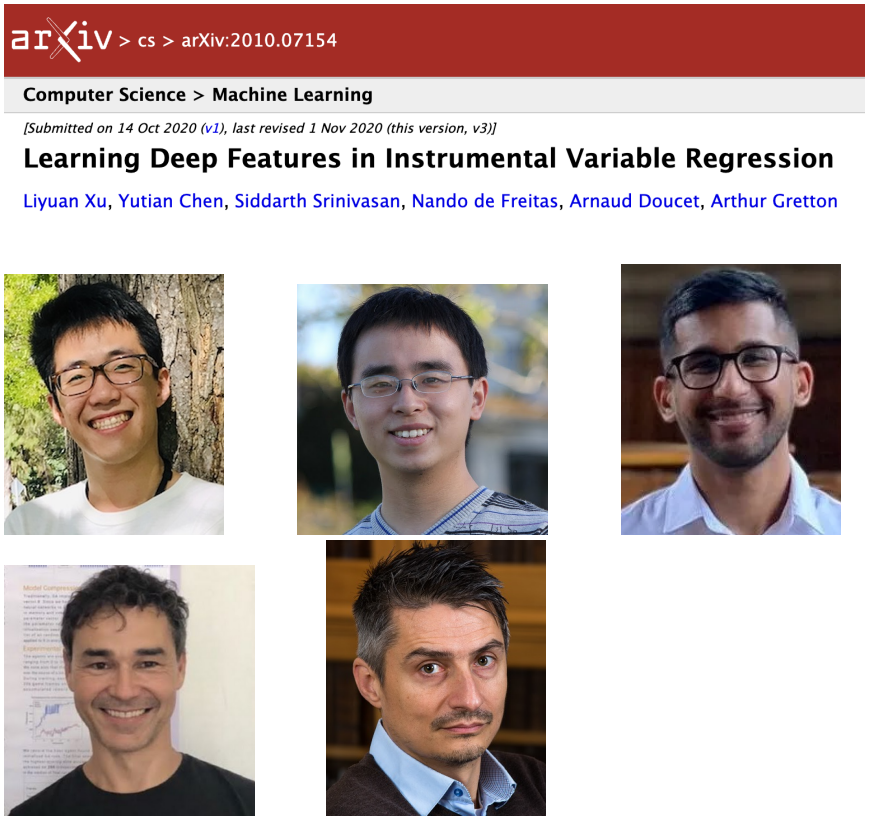
...but we don't have C_{ax}.

Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):



NN features (ICLR 2021):

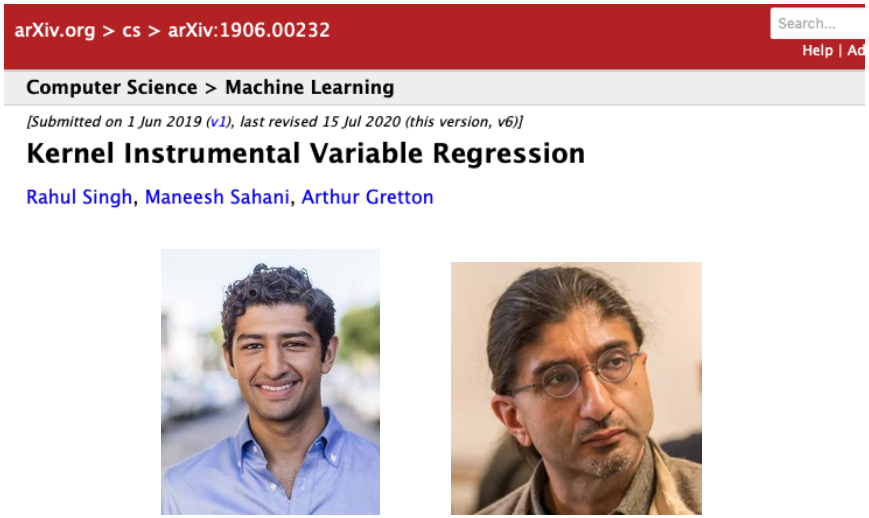


Code for NN and kernel IV methods:

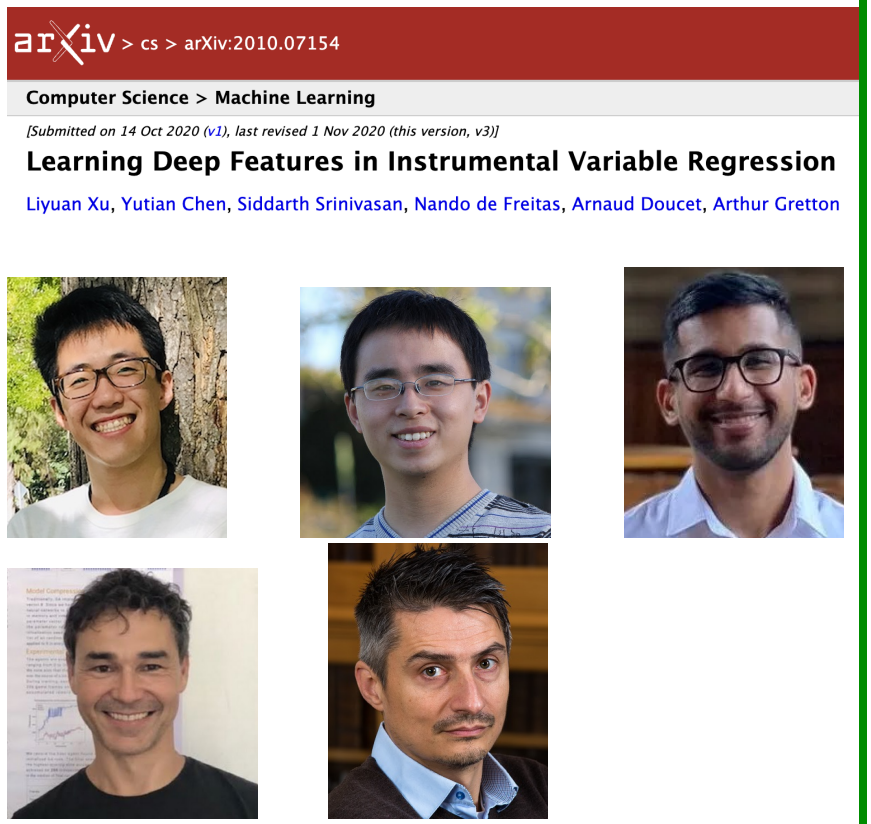
https://github.com/liyuan9988/DeepFeatureIV/

Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):



NN features (ICLR 2021):



Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[\left(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z]\right)^2\right] + \lambda_2\|\gamma\|^2$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1\|F\|_{HS}^2$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2 \|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2$$

Challenge: how to learn θ?

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2 \|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2$$

Challenge: how to learn θ?

From Stage 2 regression?

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2 \|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2$$

Challenge: how to learn θ?

From Stage 2 regression?

...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from Stage 1 regression

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2 \|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2$$

Challenge: how to learn θ?

From Stage 2 regression?
...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from Stage 1 regression
...which requires $\phi_\theta(A)$... which requires θ...

IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2 \|\gamma\|^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2$$

Challenge: how to learn θ?

From Stage 2 regression?
...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from Stage 1 regression
...which requires $\phi_\theta(A)$... which requires θ...

Use the linear final layers! (i.e. γ and F)

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

IV using neural net features

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\left[\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2\right] + \lambda_1\|F\|^2_{HS}$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

IV using neural net features

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\left[\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2\right] + \lambda_1\|F\|_{HS}^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt ϕ_θ, ϕ_ζ:

$$\hat{F}_{\theta,\zeta} = C_{AZ}(C_{ZZ} + \lambda_1 I)^{-1} \qquad C_{AZ} = \mathbb{E}[\phi_\theta(A)\phi_\zeta^\top(Z)]$$
$$C_{ZZ} = \mathbb{E}[\phi_\zeta(Z)\phi_\zeta^\top(Z)]$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

IV using neural net features

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}\left[\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2\right] + \lambda_1\|F\|_{HS}^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt ϕ_θ, ϕ_ζ:

$$\hat{F}_{\theta,\zeta} = C_{AZ}(C_{ZZ} + \lambda_1 I)^{-1} \qquad C_{AZ} = \mathbb{E}[\phi_\theta(A)\phi_\zeta^\top(Z)]$$
$$C_{ZZ} = \mathbb{E}[\phi_\zeta(Z)\phi_\zeta^\top(Z)]$$

Plug $\hat{F}_{\theta,\zeta}$ into S1 loss, take gradient steps for ζ (...but not θ...)

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable Regresion

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathcal{L}_2(\gamma, \theta) = \mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Alternative: functional bilevel optimization: Petrulionyte, Mairal, Arbel (2024)

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\begin{aligned}\mathcal{L}_2(\gamma, \theta) &= \mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2 \\ &= \mathbb{E}_{YZ}[(Y - \gamma^\top \underbrace{\hat{F}_{\theta,\zeta}\phi_\zeta(Z)}_{\text{Stage 1}})^2] + \lambda_2\|\gamma\|^2\end{aligned}$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)
Alternative: functional bilevel optimization: Petrulionyte, Mairal, Arbel (2024)

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\begin{aligned}\mathcal{L}_2(\gamma, \theta) &= \mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2 \\ &= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta}\phi_\zeta(Z))^2] + \lambda_2\|\gamma\|^2\end{aligned}$$

$\hat{\gamma}_\theta$ in closed form wrt ϕ_θ:

$$\hat{\gamma}_\theta := \widetilde{C}_{YA|Z}(\widetilde{C}_{AA|Z} + \lambda_2 I)^{-1} \qquad \widetilde{C}_{YA|Z} = \mathbb{E}\left[Y\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

$$\widetilde{C}_{AA|Z} = \mathbb{E}\left[[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Alternative: functional bilevel optimization: Petrulionyte, Mairal, Arbel (2024)

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\begin{aligned}\mathcal{L}_2(\gamma, \theta) &= \mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2 \\ &= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta}\phi_\zeta(Z))^2] + \lambda_2\|\gamma\|^2\end{aligned}$$

$\hat{\gamma}_\theta$ in closed form wrt ϕ_θ:

$$\hat{\gamma}_\theta := \widetilde{C}_{YA|Z}(\widetilde{C}_{AA|Z} + \lambda_2 I)^{-1} \qquad \widetilde{C}_{YA|Z} = \mathbb{E}\left[Y\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

$$\widetilde{C}_{AA|Z} = \mathbb{E}\left[[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

From linear final layers in Stages 1,2:

Learn $\phi_\theta(A)$ by plugging $\hat{\gamma}_\theta$ into S2 loss, taking gradient steps for θ

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Alternative: functional bilevel optimization: Petrulionyte, Mairal, Arbel (2024)

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer γ to obtain Y with RR loss:

$$\begin{aligned}\mathcal{L}_2(\gamma, \theta) &= \mathbb{E}_{YZ}\left[(Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2\right] + \lambda_2\|\gamma\|^2 \\ &= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta}\phi_\zeta(Z))^2] + \lambda_2\|\gamma\|^2\end{aligned}$$

$\hat{\gamma}_\theta$ in closed form wrt ϕ_θ:

$$\hat{\gamma}_\theta := \widetilde{C}_{YA|Z}(\widetilde{C}_{AA|Z} + \lambda_2 I)^{-1} \qquad \widetilde{C}_{YA|Z} = \mathbb{E}\left[Y\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

$$\widetilde{C}_{AA|Z} = \mathbb{E}\left[[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]\,[\hat{F}_{\theta,\zeta}\varphi_\zeta(Z)]^\top\right]$$

From linear final layers in Stages 1,2:

Learn $\phi_\theta(A)$ by plugging $\hat{\gamma}_\theta$ into S2 loss, taking gradient steps for θ

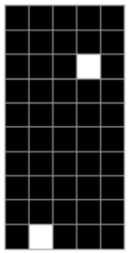
....but ζ changes with θ

...so alternate first and second stages until convergence.

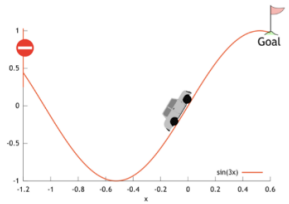
Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Alternative: functional bilevel optimization: Petrulionyte, Mairal, Arbel (2024)

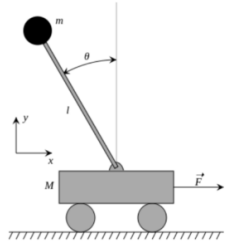
Neural IV in reinforcement learning



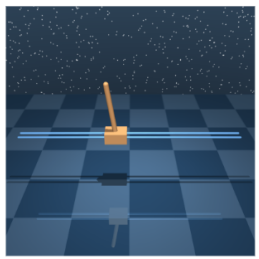
(a) Catch



(b) Mountain Car



(c) Cartpole



(a) Cartpole Swingup

(b) Cheetah Run

(c) Humanoid Run

(d) Walker Walk

Policy evaluation: want Q-value:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, A_0 = a\right]$$

for policy $\pi(A|S = s)$.

Osband et al (2019). Behaviour suite for reinforcement learning.https://github.com/deepmind/bsuite
Tassa et al. (2020). dm_control:Software and tasks for continuous control.
https://github.com/deepmind/dm_control

Application of IV: reinforcement learning

Q value is a minimizer of Bellman loss

$$\mathcal{L}_{\text{Bellman}} = \mathbb{E}_{SAR}\left[\left(R + \gamma[\mathbb{E}\left[Q^{\pi}(S', A')|S, A\right] - Q^{\pi}(S, A)\right)^2\right].$$

Corresponds to "IV-like" problem

$$\mathcal{L}_{\text{Bellman}} = \mathbb{E}_{YZ}\left[\left(Y - \mathbb{E}[f(X)|Z]\right)^2\right]$$

with

$$\begin{aligned} Y &= R, \\ X &= (S', A', S, A) \\ Z &= (S, A), \\ f_0(X) &= Q^{\pi}(s, a) - \gamma Q^{\pi}(s', a') \end{aligned}$$

RL experiments and data:

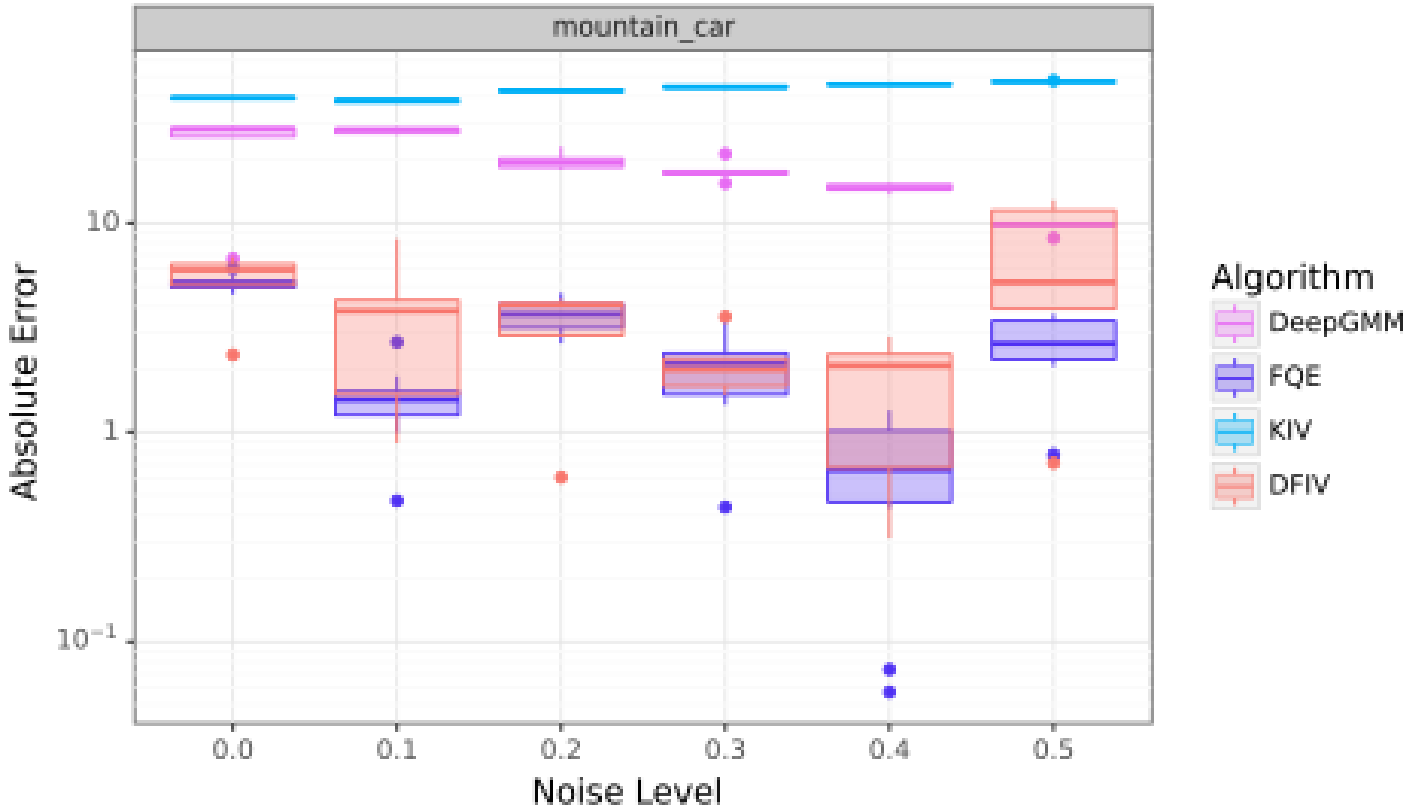
https://github.com/liyuan9988/IVOPEwithACME

Bradtke and Barto (1996). Linear least-squares algorithms for temporal difference learning.

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Chen, Xu, Gulcehre, Le Paine, G, De Freitas, Doucet (2022). On Instrumental Variable Regression for Deep Offline Policy Evaluation.

Results on mountain car problem



Good performance compared with FQE.

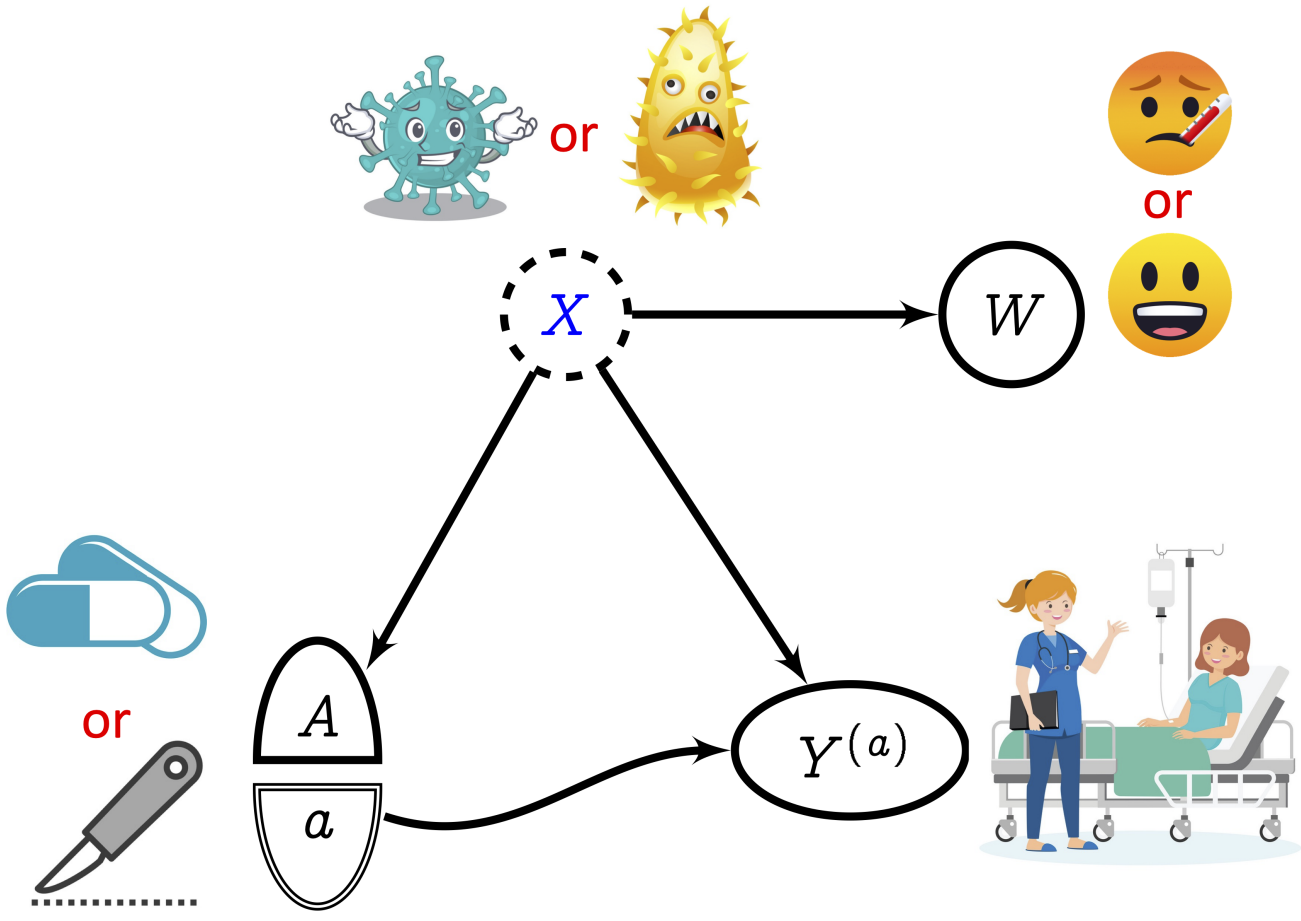
Warning: IV assumption can fail when regression underfits. See papers for details.

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Chen, Xu, Gulcehre, Le Paine, G, De Freitas, Doucet (2022). On Instrumental Variable Regression for Deep Offline Policy Evaluation.

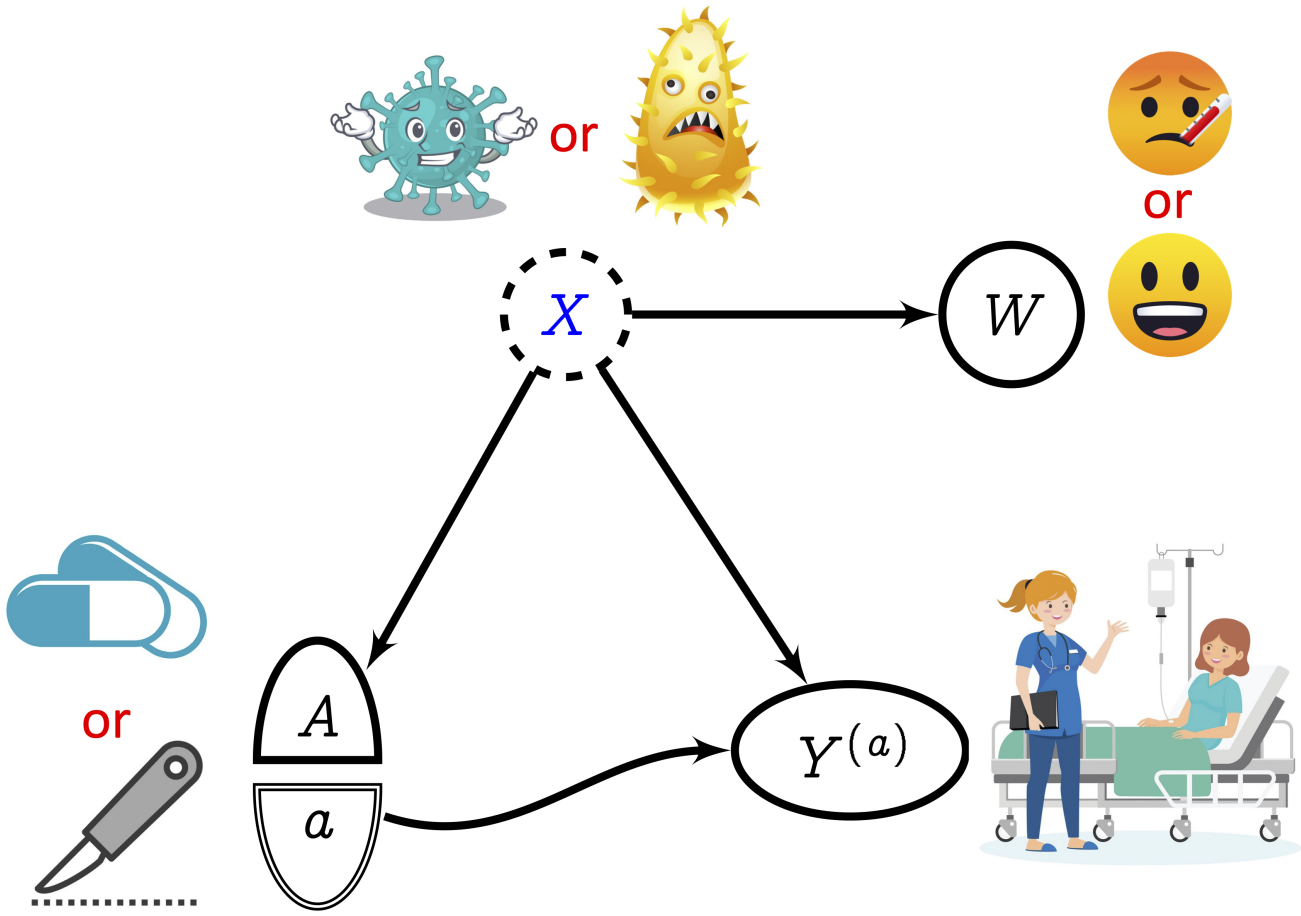
...but seriously, what if there are hidden confounders?

We record symptom W, not disease X



- $P(W = \text{fever} | X = \text{mild}) = 0.2$
- $P(W = \text{fever} | X = \text{severe}) = 0.8$

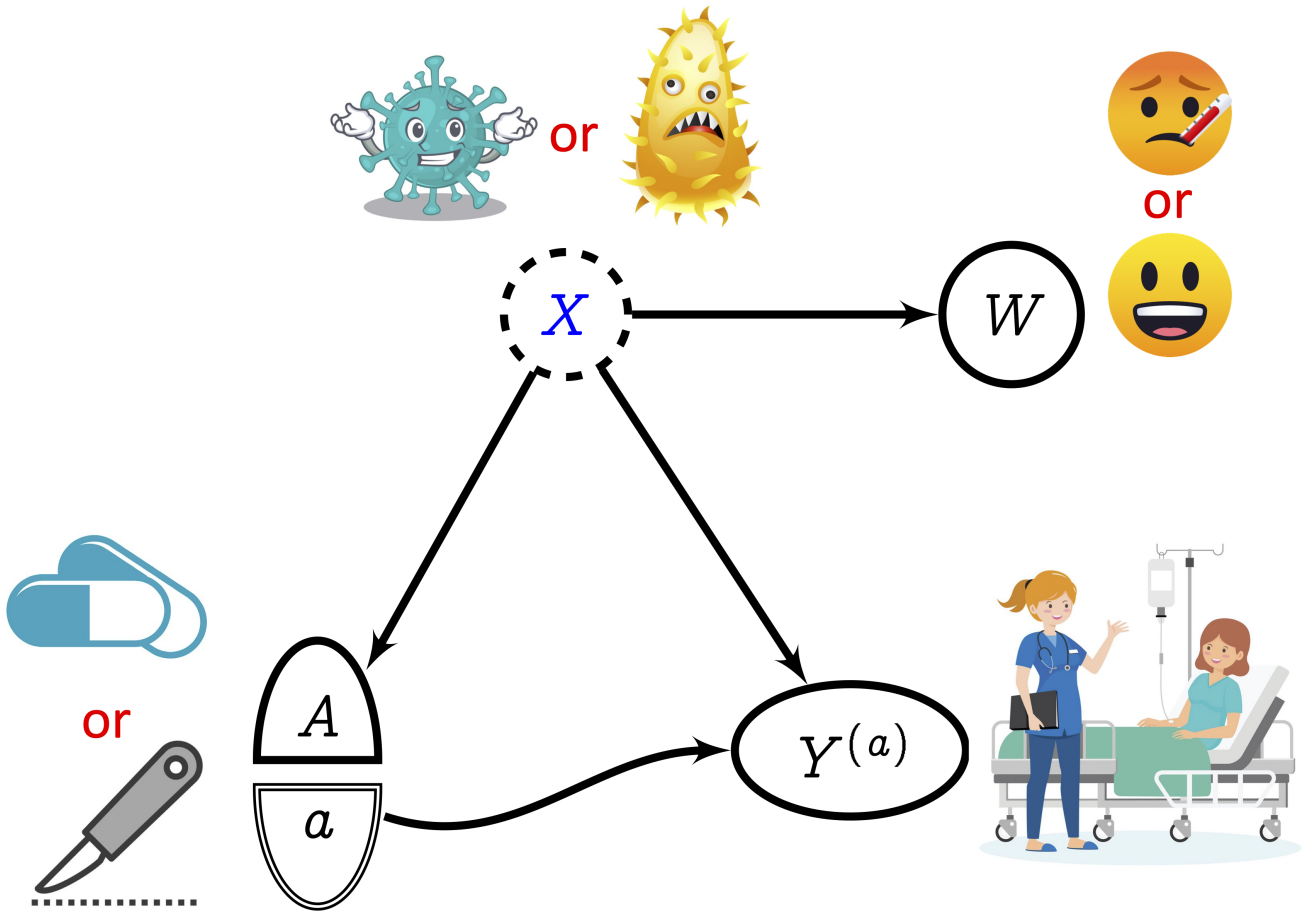
We record symptom W, not disease X



- $P(W = \text{fever} | X = \text{mild}) = 0.2$
- $P(W = \text{fever} | X = \text{severe}) = 0.8$

Could we just write: $P(Y^{(a)}) \stackrel{?}{=} \sum_{w \in \{0,1\}} \mathbb{E}[Y | a, w] p(w)$

We record symptom W, not disease X



Wrong recommendation made:

- $\sum_{w\in\{0,1\}} \mathbb{E}[\text{cured}|\text{pills}, w]p(w) = 0.8 \quad (\neq 0.64)$
- $\sum_{w\in\{0,1\}} \mathbb{E}[\text{cured}|\text{surgery}, w]p(w) = 0.73 \quad (\neq 0.75)$

Correct answer impossible without observing X

Proxy causal learning

Biometrika (2018), **105**, 4, *pp.* 987–993
Printed in Great Britain

doi: 10.1093/biomet/asy038
Advance Access publication 13 August 2018

Identifying causal effects with proxy variables of an unmeasured confounder

BY WANG MIAO

Guanghua School of Management, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, China

mwfy@pku.edu.cn

ZHI GENG

School of Mathematical Sciences, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, China

zhigeng@pku.edu.cn

AND ERIC J. TCHETGEN TCHETGEN

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

etchetge@hsph.harvard.edu

Proxy causal learning (negative controls)

Causal effect estimation, with hidden covariates X:

- Use proxy variables (negative controls)

Applications: effect of actions under

- privacy constraints (email, ads, DMA)
- data gathering constraints (edge computing)
- fundamental limitations (preferences, state of mind)

Proxy causal learning (negative controls)

Causal effect estimation, with hidden covariates X:

- Use proxy variables (negative controls)

Applications: effect of actions under

- privacy constraints (email, ads, DMA)
- data gathering constraints (edge computing)
- fundamental limitations (preferences, state of mind)

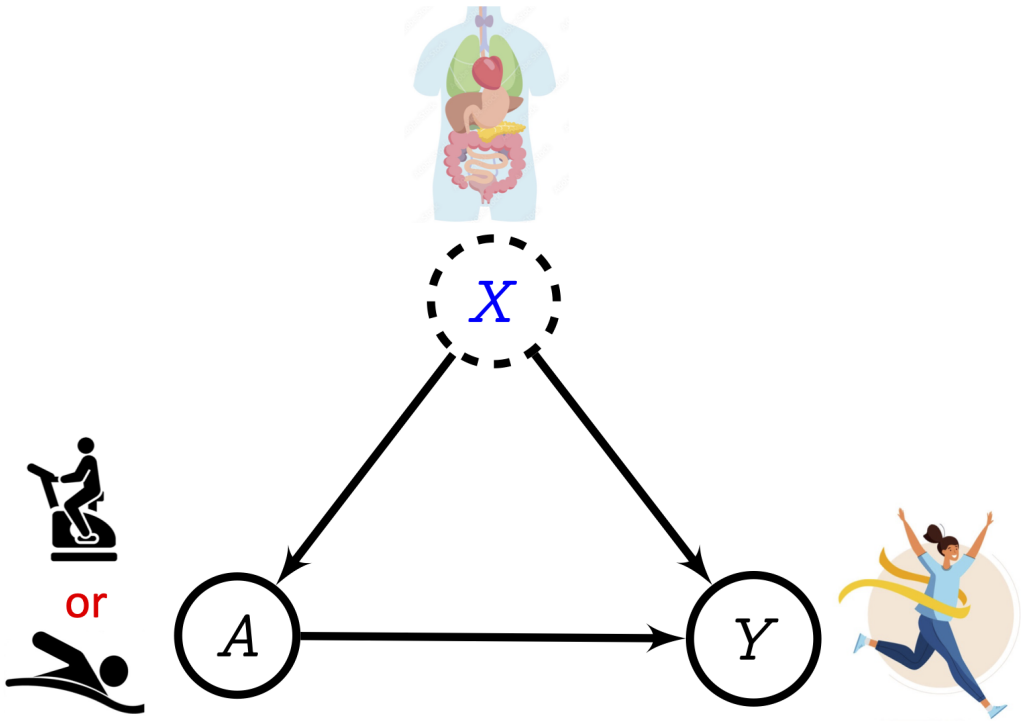
Don't ~~meet your heroes~~ model your hidden variables!

What are proxies, and when are they useful?

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal

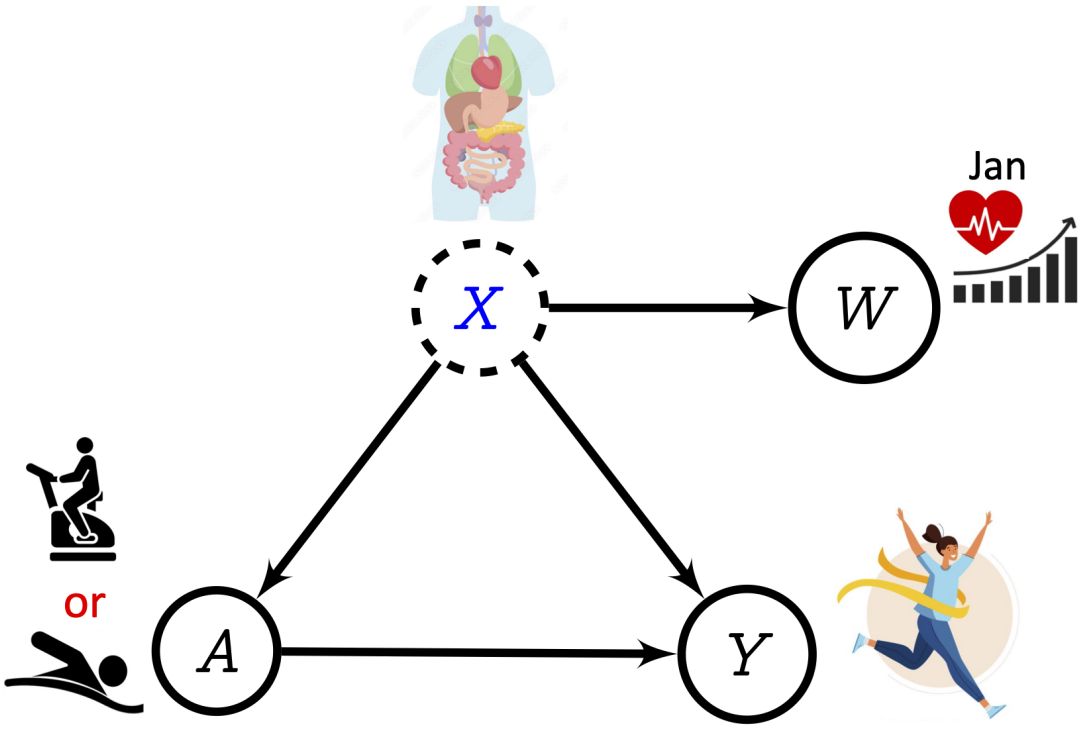


What are proxies, and when are they useful?

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A

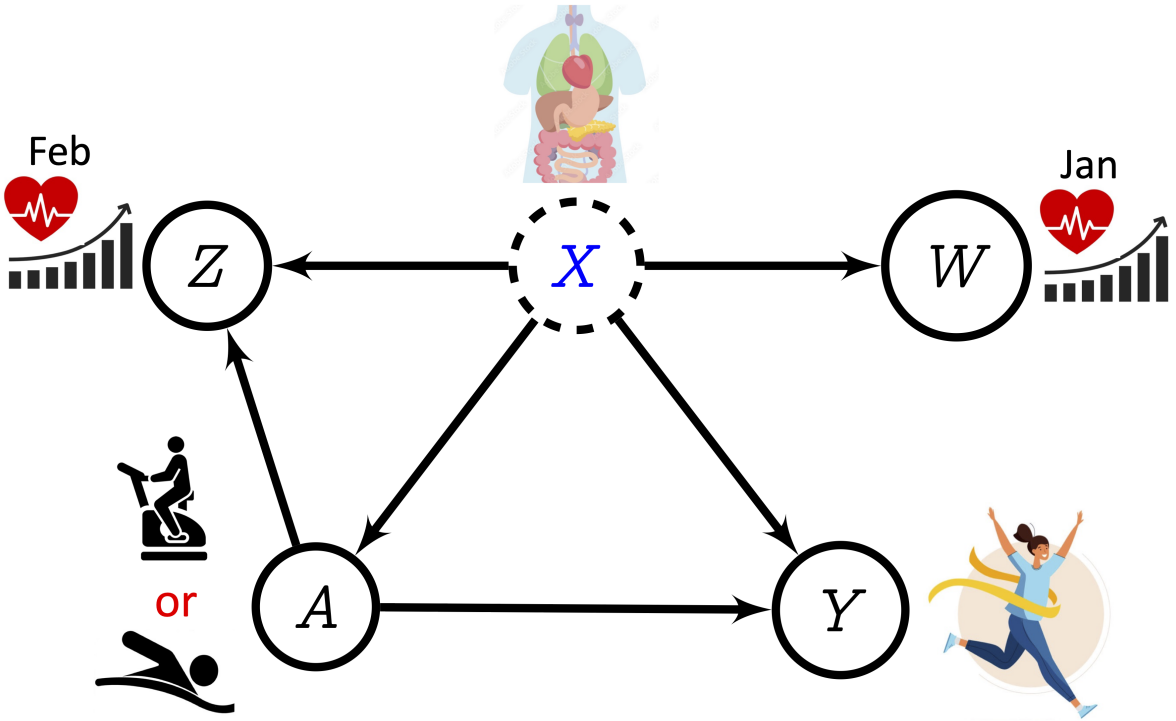


What are proxies, and when are they useful?

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A
- Z: health readings after A

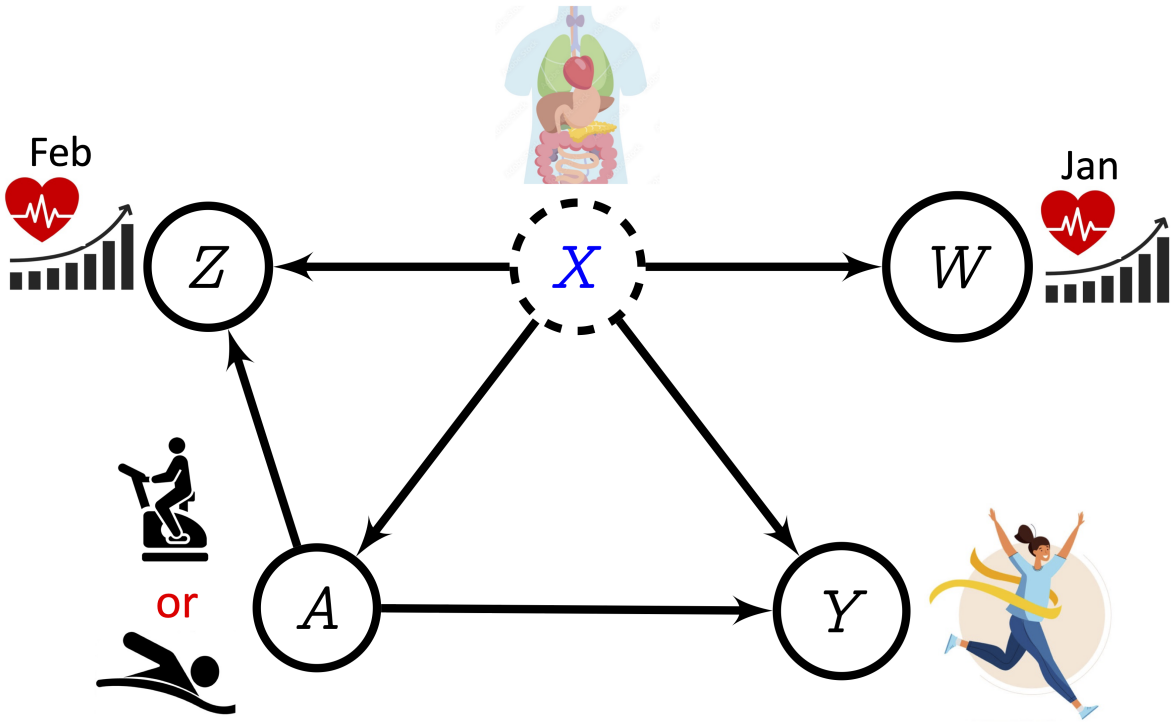


What are proxies, and when are they useful?

Unobserved X with (possibly) complex nonlinear effects on A, Y

In this example:

- X: true physical status
- A: exercise regimes
- Y: fitness goal
- W: health readings before A
- Z: health readings after A



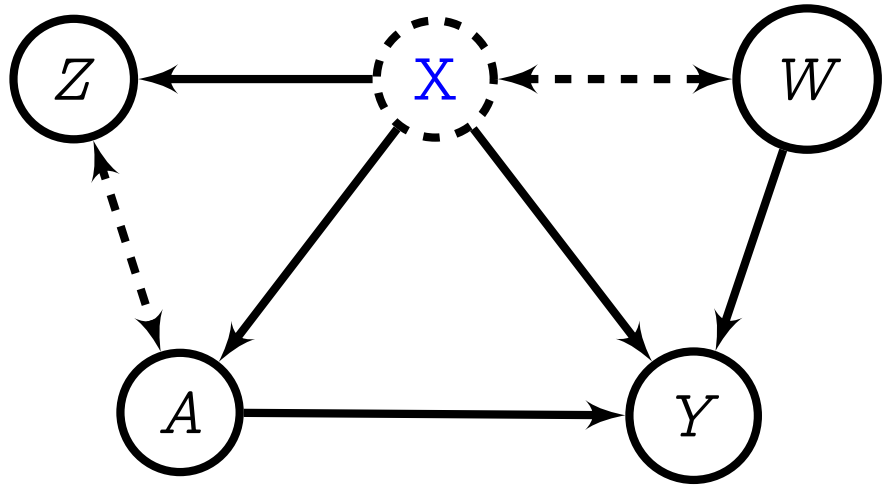
$\Longrightarrow$ Can recover $\mathbb{E}(Y^{(a)})$ from observational data

Proxy variables: general setting

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: treatment proxy
- W outcome proxy



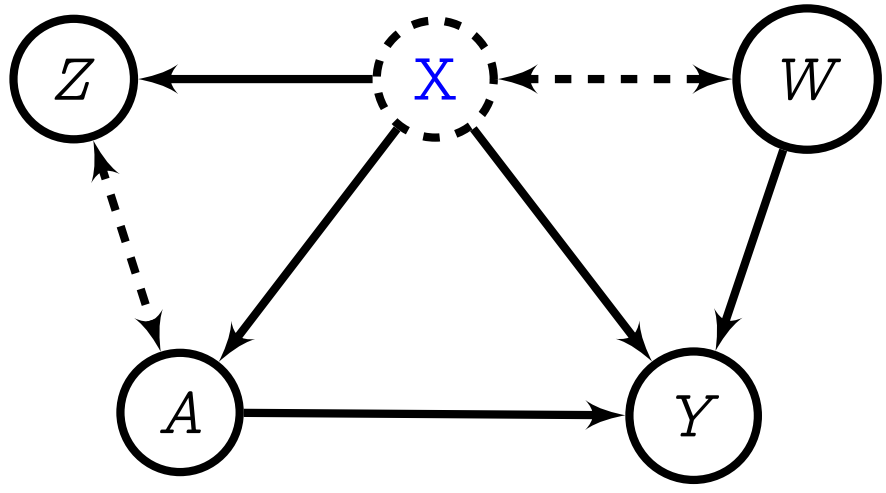
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: general setting

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: treatment proxy
- W outcome proxy



Structural assumptions:

$$W \perp\!\!\!\perp (Z, A) | X$$

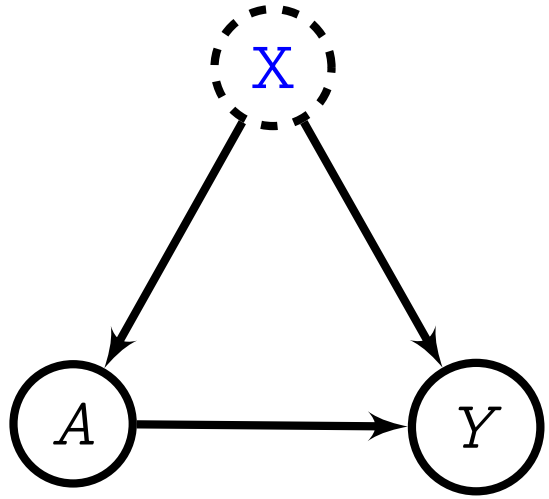
$$Y \perp\!\!\!\perp Z | (A, X)$$

Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



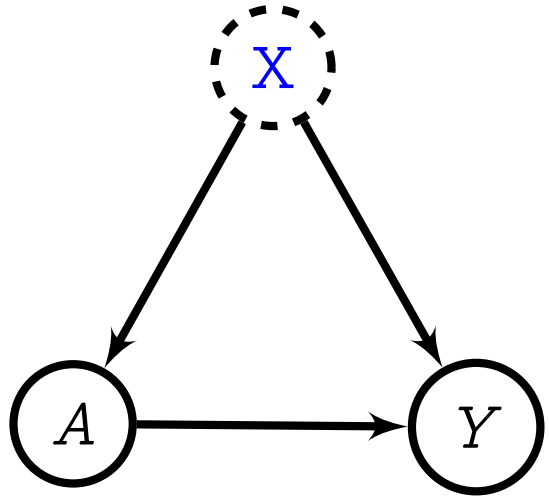
If X were observed,

$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i)$$

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



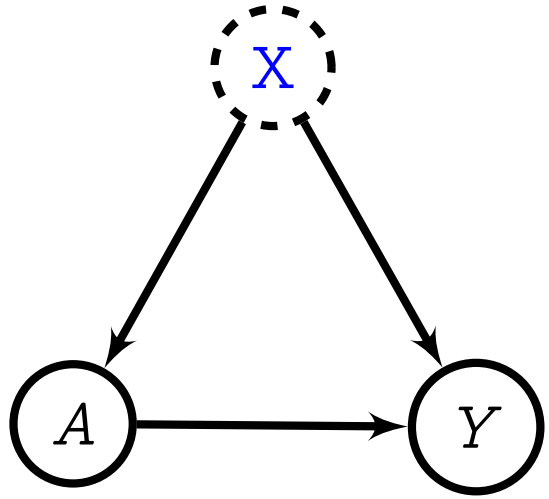
If X were observed,

$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

Why proxy variables? A simple proof

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome



If X were observed,

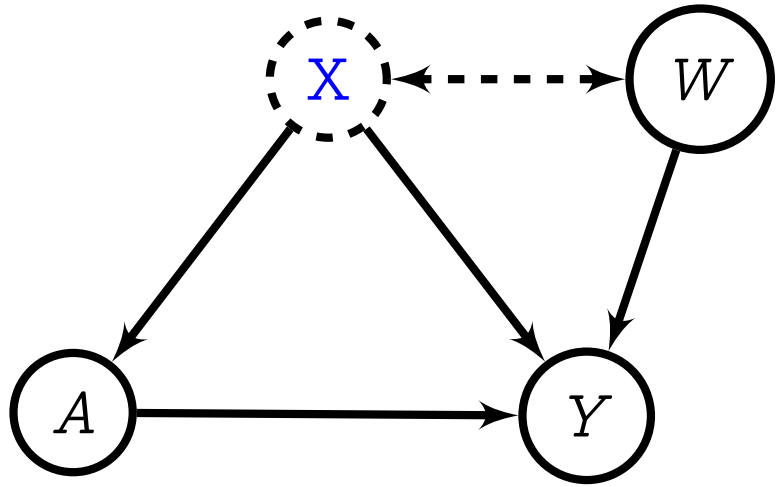
$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

Goal: "get rid of the blue" X

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



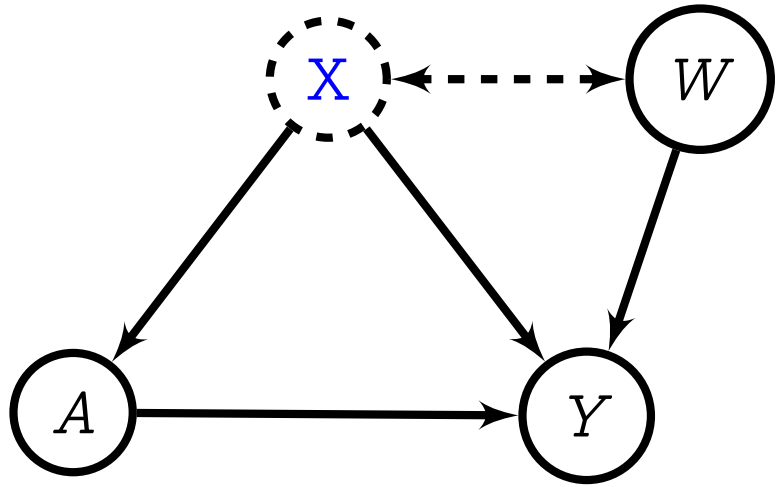
For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

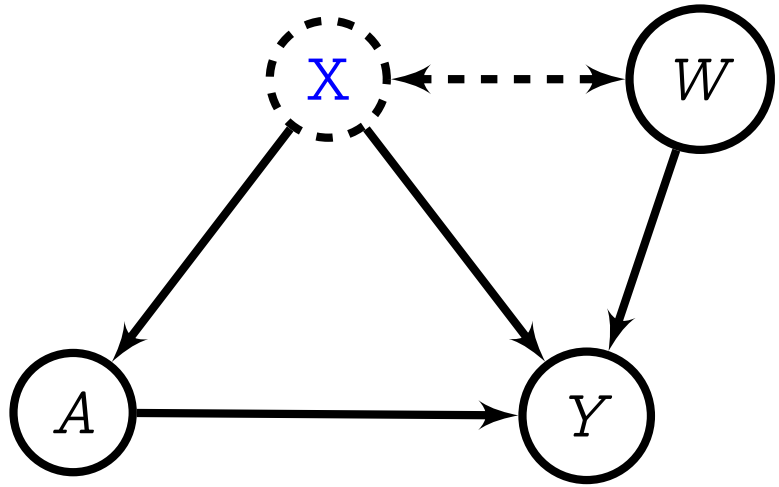
.....then

$$P(Y^{(a)}) = P(Y|X,a)P(X)$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

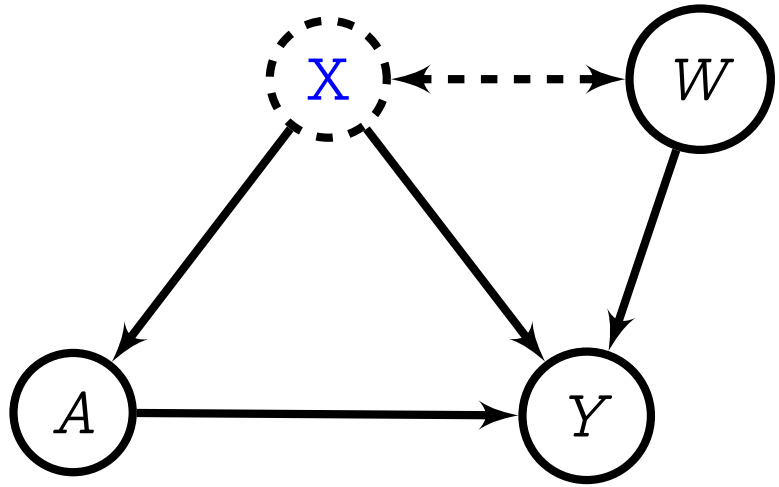
.....then

$$\begin{aligned} P(Y^{(a)}) &= P(Y|X,a)P(X) \\ &= H_{w,a}P(W|X)P(X) \end{aligned}$$

...add the outcome proxy W

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- W: outcome proxy



For each a, if we could solve:

$$\underbrace{P(Y|X,a)}_{d_y \times d_x} = \underbrace{H_{w,a}}_{d_y \times d_w} \underbrace{P(W|X)}_{d_w \times d_x}$$

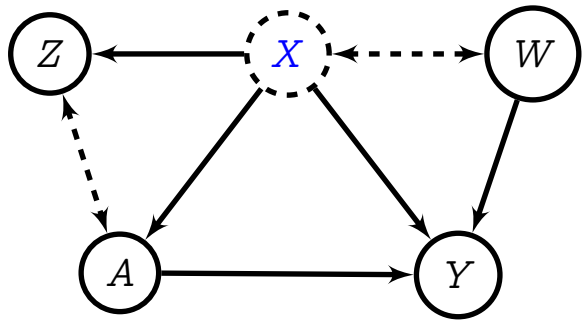
.....then

$$\begin{aligned} P(Y^{(a)}) &= P(Y|X,a)P(X) \\ &= H_{w,a}P(W|X)P(X) \\ &= H_{w,a}P(W) \end{aligned}$$

...now project onto $p(X|Z,a)$

From last slide,

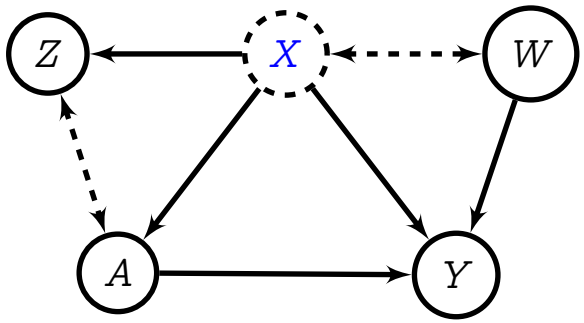
$$P(Y|X,a) = H_{w,a} P(W|X)$$



...now project onto $p(X|Z,a)$

From last slide,

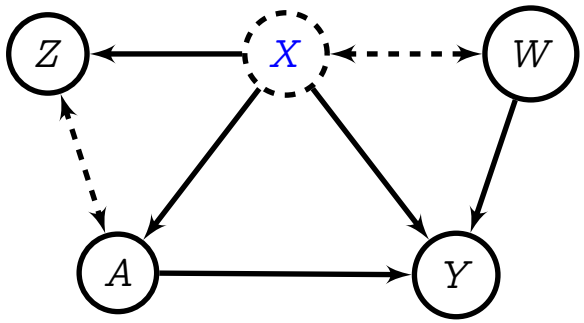
$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X) \underbrace{p(X|Z,a)}_{d_x \times d_z}$$



...now project onto $p(X|Z,a)$

From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X) \underbrace{p(X|Z,a)}_{d_x \times d_z}$$



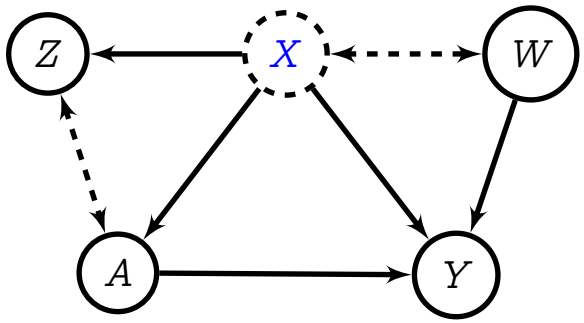
Because $W \perp\!\!\!\perp (Z, A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

...now project onto $p(X|Z,a)$

From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X)\underbrace{p(X|Z,a)}_{d_x \times d_z}$$



Because $W \perp\!\!\!\perp (Z, A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

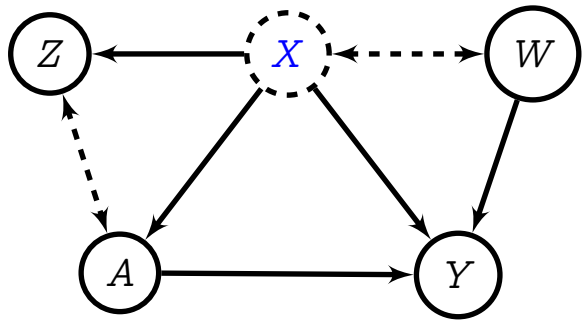
Because $Y \perp\!\!\!\perp Z|(A,X)$,

$$P(Y|X,a)p(X|Z,a) = P(Y|Z,a)$$

...now project onto $p(X|Z,a)$

From last slide,

$$P(Y|X,a)\underbrace{p(X|Z,a)}_{d_x \times d_z} = H_{w,a} P(W|X) \underbrace{p(X|Z,a)}_{d_x \times d_z}$$



Because $W \perp\!\!\!\perp (Z, A)|X$,

$$P(W|X)p(X|Z,a) = P(W|Z,a)$$

Because $Y \perp\!\!\!\perp Z|(A,X)$,

$$P(Y|X,a)p(X|Z,a) = P(Y|Z,a)$$

Solve for $H_{w,a}$:

$$P(Y|Z,a) = H_{w,a} P(W|Z,a)$$

Everything observed!

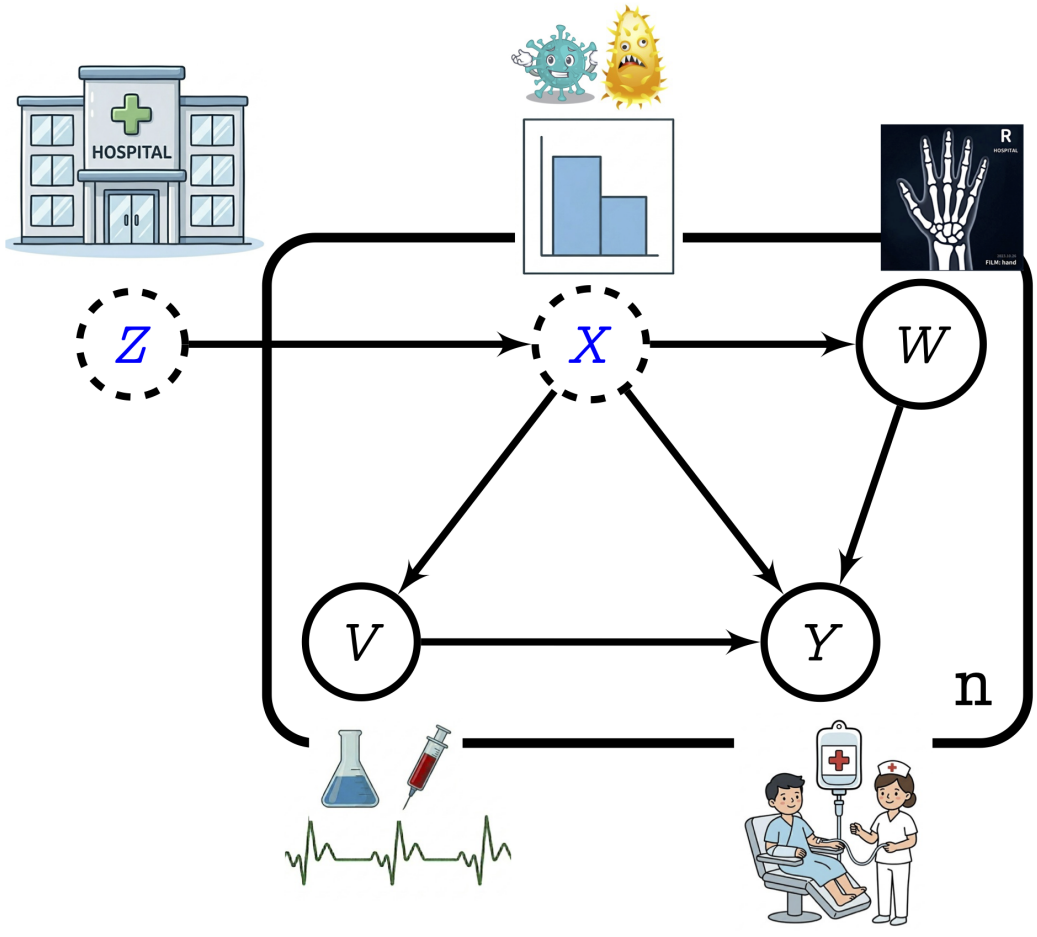
Domain shift: the blessings of multiple domains

Outcome bridge for domain shift

"The blessings of multiple domains"

In this example:

- X: which disease
- V: blood tests, ECG
- W: x-rays
- Y: diagnosis
- Z: domain (P_X param.)
- $n \in \{1 \dots D+1\}$



arXiv > cs > arXiv:2403.07442

Computer Science > Machine Learning

[Submitted on 12 Mar 2024]

Proxy Methods for Domain Adaptation

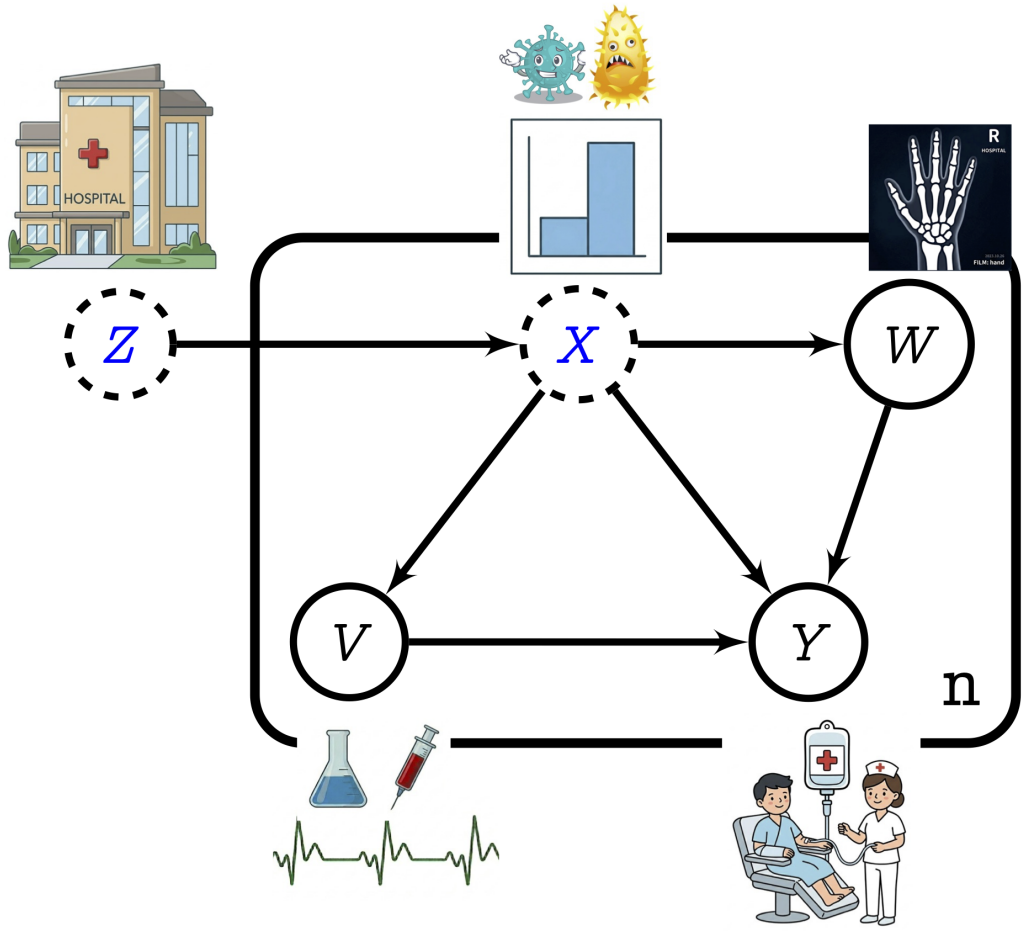
Katherine Tsai, Stephen R. Pfohl, Olawale Salaudeen, Nicole Chiou, Matt J. Kusner, Alexander D'Amour, Sanmi Koyejo, Arthur Gretton

Outcome bridge for domain shift

"The blessings of multiple domains"

In this example:

- X: which disease
- V: blood tests, ECG
- W: x-rays
- Y: diagnosis
- Z: domain (P_X param.)
- $n \in \{1 \dots D+1\}$



arXiv > cs > arXiv:2403.07442

Computer Science > Machine Learning

[Submitted on 12 Mar 2024]

Proxy Methods for Domain Adaptation

Katherine Tsai, Stephen R. Pfohl, Olawale Salaudeen, Nicole Chiou, Matt J. Kusner, Alexander D'Amour, Sanmi Koyejo, Arthur Gretton

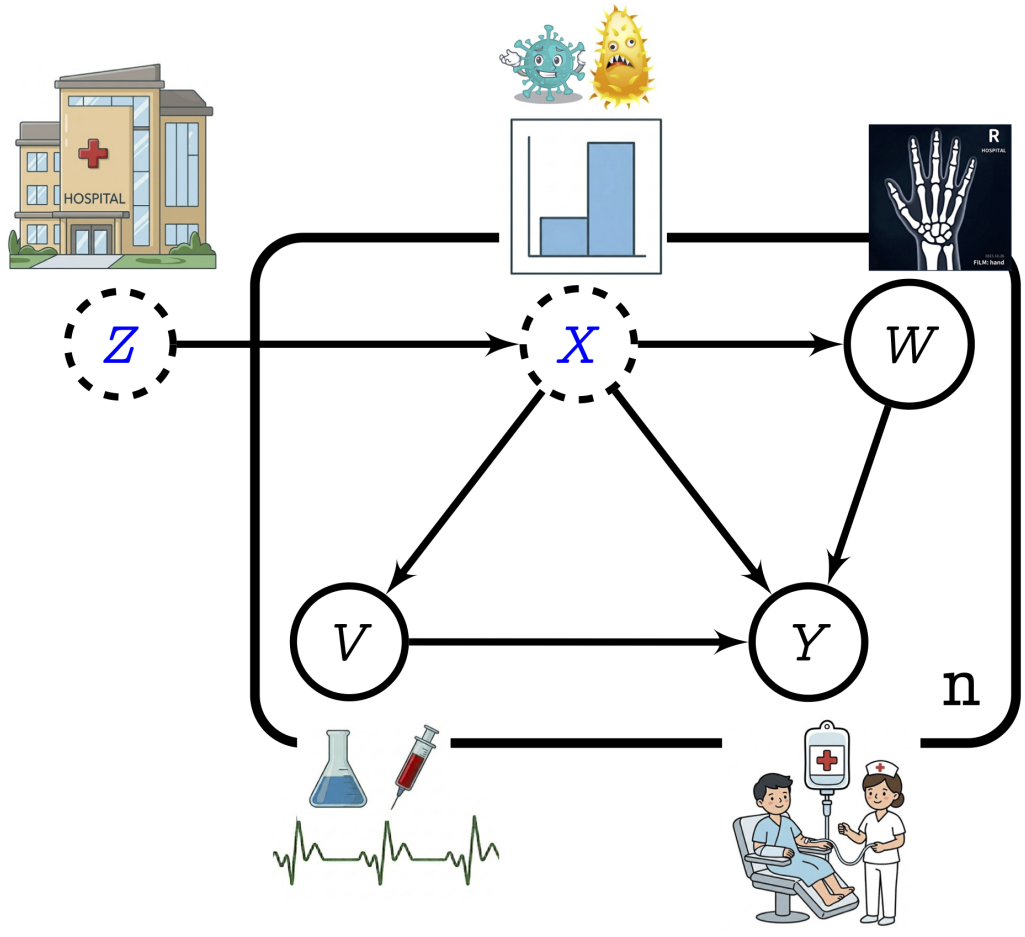
Outcome bridge for domain shift

"The blessings of multiple domains"

In this example:

- X: which disease
- V: blood tests, ECG
- W: x-rays
- Y: diagnosis
- Z: domain (P_X param.)
- $n \in \{1 \dots D+1\}$

Goal: prediction on new (*) domain D+1 via outcome bridge



arXiv > cs > arXiv:2403.07442

Computer Science > Machine Learning

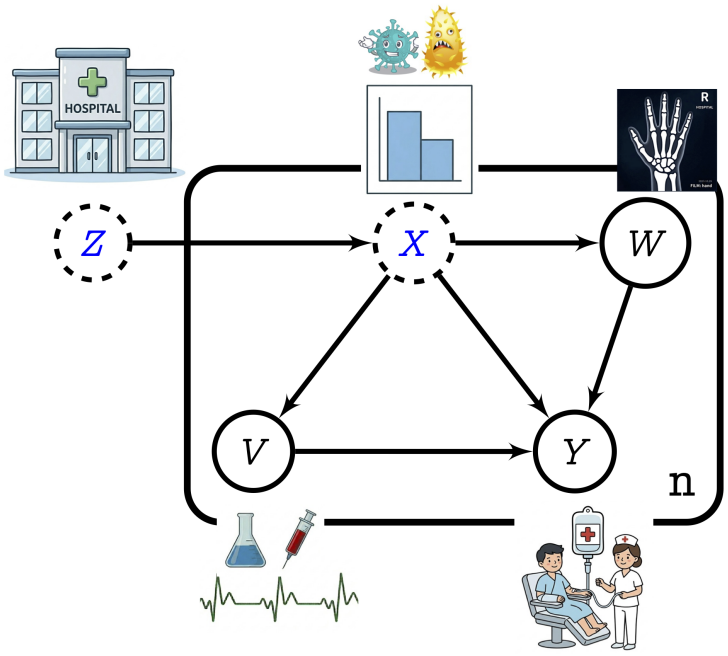
[Submitted on 12 Mar 2024]

Proxy Methods for Domain Adaptation

Katherine Tsai, Stephen R. Pfohl, Olawale Salaudeen, Nicole Chiou, Matt J. Kusner, Alexander D'Amour, Sanmi Koyejo, Arthur Gretton

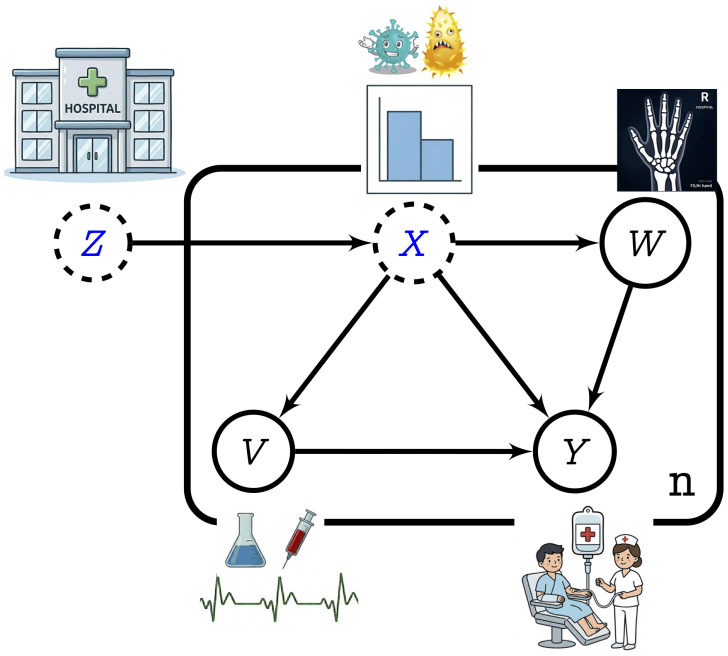
Proof (1)

- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



Proof (1)

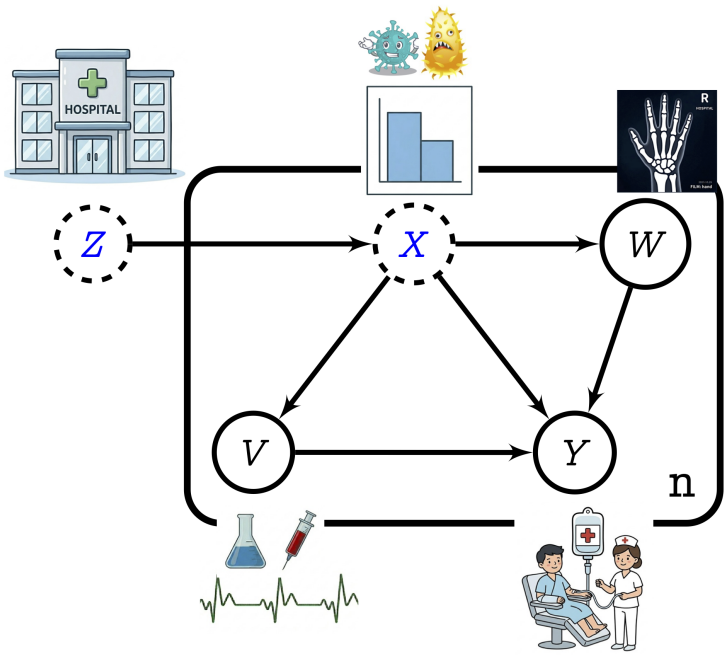
- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

Proof (1)

- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



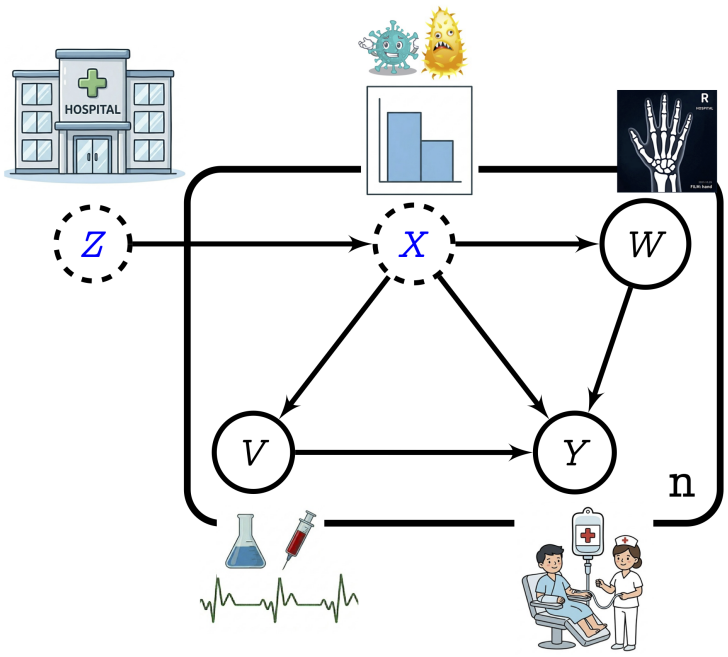
Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

For a new domain $(*)$ given outcome bridge $M_{w,v}$ solving

$$P(Y|X, v) = M_{w,v} P(W|X)$$

Proof (1)

- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

For a new domain $(*)$ given outcome bridge $M_{w,v}$ solving

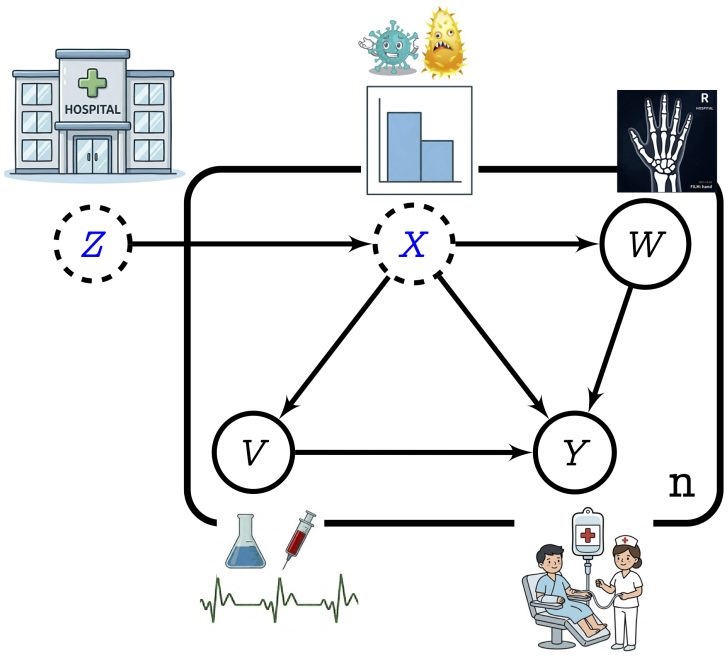
$$P(Y|X, v) = M_{w,v} P(W|X)$$

....then

$$P^{(*)}(Y|v) = P(Y|X, v) P^{(*)}(X|v)$$

Proof (1)

- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

For a new domain $(*)$ given outcome bridge $M_{w,v}$ solving

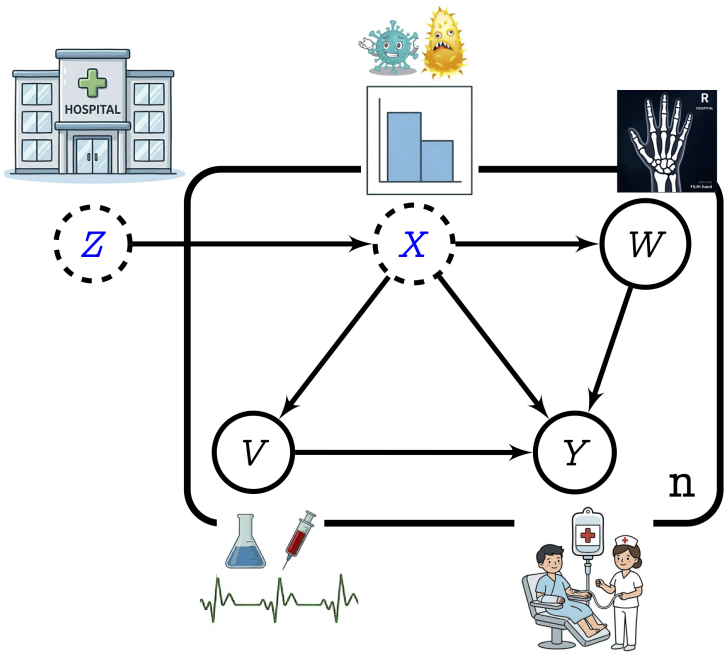
$$P(Y|X, v) = M_{w,v} P(W|X)$$

....then

$$\begin{aligned} P^{(*)}(Y|v) &= P(Y|X, v) P^{(*)}(X|v) \\ &= M_{w,v} P(W|X) P^{(*)}(X|v) \end{aligned}$$

Proof (1)

- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

For a new domain $(*)$ given outcome bridge $M_{w,v}$ solving

$$P(Y|X, v) = M_{w,v} P(W|X)$$

....then

$$\begin{aligned} P^{(*)}(Y|v) &= P(Y|X, v) P^{(*)}(X|v) \\ &= M_{w,v} P(W|X) P^{(*)}(X|v) \\ &= M_{w,v} P^{(*)}(W|v) \end{aligned}$$

Proof (2) - how to solve for outcome bridge?

Get rid of the blue!

For each v, define matrix of per-domain probabilities:

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \left[\begin{array}{ccc} P^{(1)}(X|v) & \ldots & P^{(D)}(X|v) \end{array} \right]$$

Proof (2) - how to solve for outcome bridge?

Get rid of the blue!

For each v, define matrix of per-domain probabilities:

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \left[\begin{array}{ccc} P^{(1)}(X|v) & \ldots & P^{(D)}(X|v) \end{array} \right]$$

Then

$$\begin{aligned} P^{(1..D)}(W|v) &:= \left[\begin{array}{ccc} P^{(1)}(W|v) & \ldots & P^{(D)}(W|v) \end{array} \right] \\ &= P(W|X)P^{(1..D)}(X|v) \end{aligned}$$

Proof (2) - how to solve for outcome bridge?

Get rid of the blue!

For each v, define matrix of per-domain probabilities:

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \left[\begin{array}{ccc} P^{(1)}(X|v) & \ldots & P^{(D)}(X|v) \end{array} \right]$$

Then

$$\begin{aligned} P^{(1..D)}(W|v) &:= \left[\begin{array}{ccc} P^{(1)}(W|v) & \ldots & P^{(D)}(W|v) \end{array} \right] \\ &= P(W|X)P^{(1..D)}(X|v) \end{aligned}$$

Similarly

$$\begin{aligned} P^{(1..D)}(Y|v) &:= \left[\begin{array}{ccc} P^{(1)}(Y|v) & \ldots & P^{(D)}(Y|v) \end{array} \right] \\ &= P(Y|X,v)P^{(1..D)}(X|v) \end{aligned}$$

Proof (2) - how to solve for outcome bridge?

Get rid of the blue!

For each v, define matrix of per-domain probabilities:

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \left[\begin{array}{ccc} P^{(1)}(X|v) & \ldots & P^{(D)}(X|v) \end{array} \right]$$

Then

$$P^{(1..D)}(W|v) := \left[\begin{array}{ccc} P^{(1)}(W|v) & \ldots & P^{(D)}(W|v) \end{array} \right]$$
$$= P(W|X)P^{(1..D)}(X|v)$$

Similarly

$$P^{(1..D)}(Y|v) := \left[\begin{array}{ccc} P^{(1)}(Y|v) & \ldots & P^{(D)}(Y|v) \end{array} \right]$$
$$= P(Y|X,v)P^{(1..D)}(X|v)$$

Then

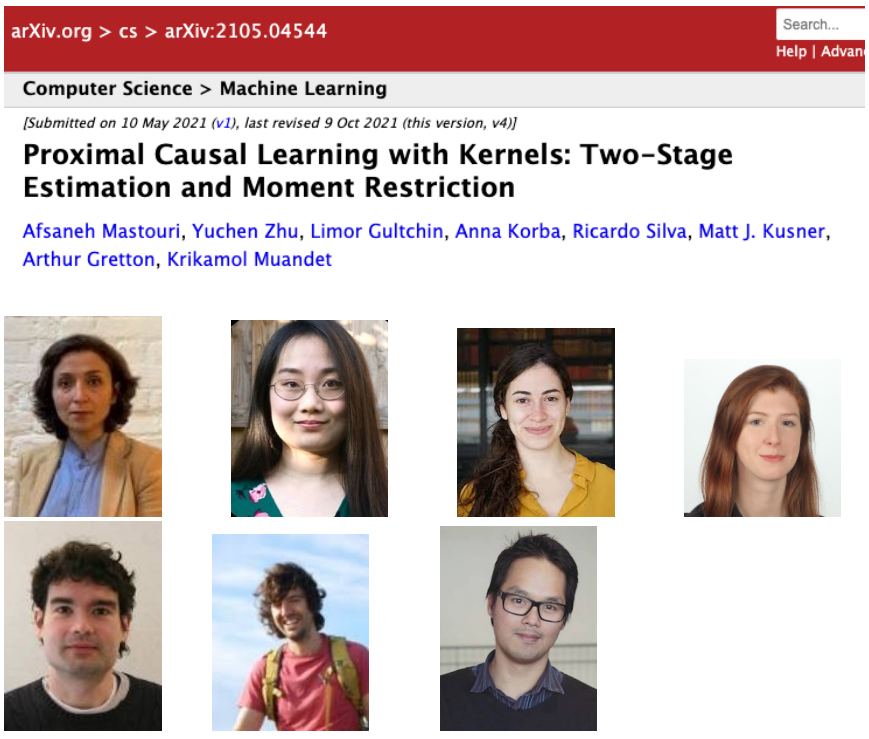
$$P^{(1..D)}(Y|v) = M_{w,v}P^{(1..D)}(W|v)$$

Solve for $M_{w,v}$ when $\text{rank}(P^{(1..D)}(W|v)) \geq d_x$.

Proxy/Negative Control Methods in the Real World

Outcome bridge and proxy variables

Kernel features (ICML 2021):



NN features (NeurIPS 2021):

Code for NN and kernel proxy methods:

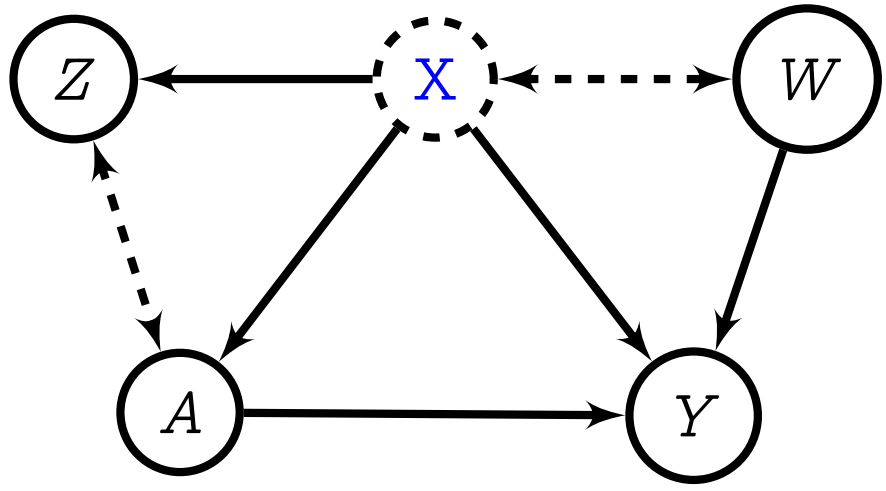
https://github.com/liyuan9988/DeepFeatureProxyVariable/

Reminder: proxy variables, finite setting

Unobserved X with (possibly) complex nonlinear effects on A, Y

The definitions are:

- X: unobserved confounder.
- A: treatment
- Y: outcome
- Z: treatment proxy
- W outcome proxy



We can obtain

$$P(Y^{(a)}) = P(Y|X,a)P(X)$$
$$= H_{w,a}P(W)$$

by solving

$$P(Y|Z,a) = H_{w,a}P(W|Z,a)$$

Proxy relation, general domains

If X were observed, we would write (dose-response curve)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a,x)p(x)dx.$$

....but we do not observe X.

Proxy relation, general domains

If X were observed, we would write (dose-response curve)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a,x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

- "Primary" $\mathbb{E}(Y|a,z)$, "secondary" $\mathbb{E}_{W|a,z}$ linked by h_y
- All variables observed, X not seen *or modeled*.

Fredholm equation of first kind. Link existence requires $\diamondsuit$, identification of ATE requires $\triangle$ (and further technical assumptions) [XKG: Asspumption 2, Prop. 1,Corr. 1; Deaner]

$$\mathbb{E}[f(X)|A=a, Z=z] = 0,\ \forall(z,a) \iff f(X) = 0,\ \mathbb{P}_X \text{ a.s.} \quad \triangle$$

$$\mathbb{E}[f(X)|A=a, W=w] = 0,\ \forall(w,a) \iff f(X) = 0,\ \mathbb{P}_X \text{ a.s.} \quad \diamondsuit$$

Proxy relation, general domains

If X were observed, we would write (dose-response curve)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a,x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

- "Primary" $\mathbb{E}(Y|a,z)$, "secondary" $\mathbb{E}_{W|a,z}$ linked by h_y
- All variables observed, X not seen *or modeled*.

Dose-response curve via $p(w)$:

$$\mathbb{E}(Y^{(a)}) = \int_w h_y(a,w)p(w)dw$$

Proxy relation, general domains

If X were observed, we would write (dose-response curve)

$$\mathbb{E}(Y^{(a)}) = \int_x \mathbb{E}(Y|a,x)p(x)dx.$$

....but we do not observe X.

Main theorem: Assume we solved for link function:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

- "Primary" $\mathbb{E}(Y|a,z)$, "secondary" $\mathbb{E}_{W|a,z}$ linked by h_y
- All variables observed, X not seen *or modeled*.

Dose-response curve via $p(w)$:

$$\mathbb{E}(Y^{(a)}) = \int_w h_y(a,w)p(w)dw$$

Challenge: need a loss function for h_y

Primary loss function for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Primary loss function:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2$$

Why?

Deaner (2021).

Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).

Xu, Kanagawa, G. (2021).

Primary loss function for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

Primary loss function:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W,A) \right)^2$$

Why?

$f^*(a,z) = \mathbb{E}(Y|a,z)$ solves

$$\operatorname{argmin}_f \mathbb{E}_{Y,A,Z} \left(Y - f(A,Z) \right)^2$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

Primary loss function for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Primary loss function:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2$$

Why?

$f^*(a,z) = \mathbb{E}(Y|a,z)$ solves

$$\operatorname{argmin}_f \mathbb{E}_{Y,A,Z} \left(Y - f(A,Z) \right)^2$$

...and by the proxy model above,

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN for link $h_y(a, w)$

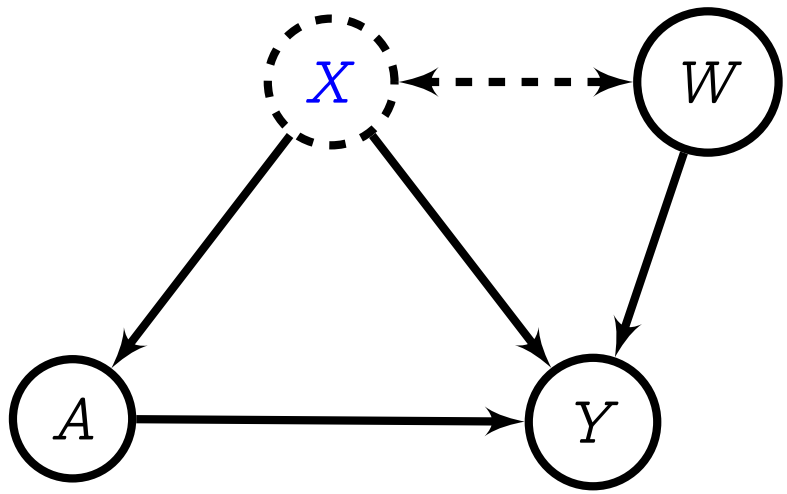
The link function is a function of two arguments

$$h_y(a, w) = \gamma^\top [\varphi_\theta(w) \otimes \varphi_\xi(a)] = \gamma^\top \begin{bmatrix} \varphi_{\theta,1}(w)\varphi_{\xi,1}(a) \\ \varphi_{\theta,1}(w)\varphi_{\xi,2}(a) \\ \vdots \\ \varphi_{\theta,2}(w)\varphi_{\xi,1}(a) \\ \vdots \end{bmatrix}$$

Assume we have:

- output proxy NN features $\varphi_\theta(w)$
- treatment NN features $\varphi_\xi(a)$
- linear final layer γ

(argument of feature map indicates feature space)



NN for link $h_y(a, w)$

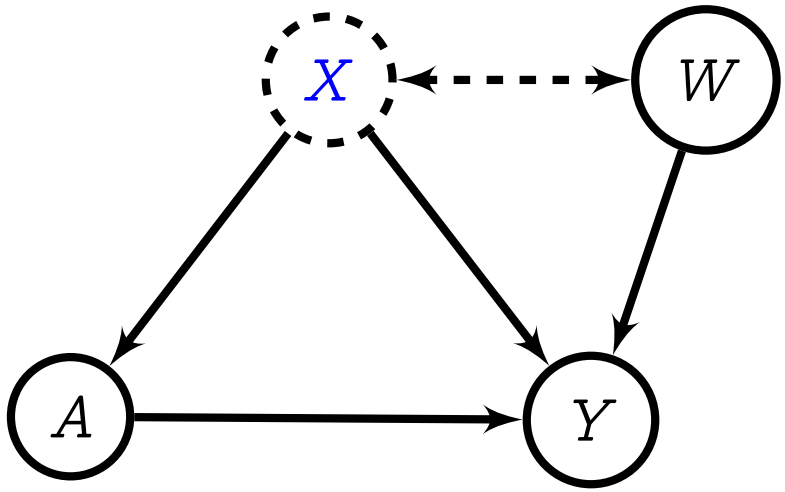
The link function is a function of two arguments

$$h_y(a, w) = \gamma^\top [\varphi_\theta(w) \otimes \varphi_\xi(a)]$$

Assume we have:

- output proxy NN features $\varphi_\theta(w)$
- treatment NN features $\varphi_\xi(a)$
- linear final layer γ

(argument of feature map indicates feature space)



Questions:

- Why feature map $\varphi_\theta(w) \otimes \varphi_\xi(a)$?
- Why final linear layer γ?

Both are necessary (next slide)!

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \Big(Y - \mathbb{E}_{W|A,Z} h_y(W,A) \Big)^2 + \lambda_2 \|\gamma\|^2$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z}\Big(Y - \mathbb{E}_{W|A,Z} h_y(W,A)\Big)^2 + \lambda_2 \|\gamma\|^2$$

How to get conditional expectation $\mathbb{E}_{W|a,z} h_y(W,a)$?

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \right)^2 + \lambda_2 \|\gamma\|^2$$

Recall link function

$$h_y(W, a) = \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(a)) \right]$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z}\Big(Y - \mathbb{E}_{W|A,Z} h_y(W,A)\Big)^2 + \lambda_2 \|\gamma\|^2$$

Recall link function

$$\mathbb{E}_{W|a,z}\ h_y(W,a) = \mathbb{E}_{W|a,z}\ \Big[\gamma^\top(\varphi_\theta(W) \otimes \varphi_\xi(a))\Big]$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \Big(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \Big)^2 + \lambda_2 \|\gamma\|^2$$

Recall link function

$$\begin{aligned} \mathbb{E}_{W|a,z}\ h_y(W, a) &= \mathbb{E}_{W|a,z} \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(a)) \right] \\ &= \gamma^\top \Big(\underbrace{\mathbb{E}_{W|a,z} [\varphi_\theta(W)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a) \Big) \end{aligned}$$

(this is why linear γ and feature map $\varphi_\theta(w) \otimes \varphi_\xi(a)$)

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

Primary regression:

$$\hat{h}_y = \arg\min_{h_y} \mathbb{E}_{Y,A,Z} \Big(Y - \mathbb{E}_{W|A,Z} h_y(W, A) \Big)^2 + \lambda_2 \|\gamma\|^2$$

Recall link function

$$\begin{aligned} \mathbb{E}_{W|a,z}\ h_y(W, a) &= \mathbb{E}_{W|a,z} \left[\gamma^\top (\varphi_\theta(W) \otimes \varphi_\xi(a)) \right] \\ &= \gamma^\top \Big(\underbrace{\mathbb{E}_{W|a,z} [\varphi_\theta(W)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a) \Big) \end{aligned}$$

Ridge regression (again!)

$$\mathbb{E}_{W|a,z} \varphi_\theta(W) = \hat{F}_{\theta,\zeta} \varphi_\zeta(a, z)$$

Deaner (2021).
Mastouri, Zhu, Gultchin, Korba, Silva, Kusner, G., Muandet (2021).
Xu, Kanagawa, G. (2021).

NN ridge regression for $\mathbb{E}_{W|a,z}\varphi_\theta(W)$

Secondary regression: learn NN features $\varphi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}_{W|a,z}\varphi_\theta(W) = \hat{F}_{\theta,\zeta}\varphi_\zeta(a,z)$$

with RR loss

$$\mathbb{E}_{W,A,Z}\|\varphi_\theta(W) - F\varphi_\zeta(A,Z)\|^2 + \lambda_1\|F\|^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt $\varphi_\theta, \varphi_\zeta$.

Xu, Kanagawa, G. (2021).

NN ridge regression for $\mathbb{E}_{W|a,z}\varphi_\theta(W)$

Secondary regression: learn NN features $\varphi_\zeta(Z)$ and linear layer F:

$$\mathbb{E}_{W|a,z}\varphi_\theta(W) = \hat{F}_{\theta,\zeta}\varphi_\zeta(a,z)$$

with RR loss

$$\mathbb{E}_{W,A,Z}\|\varphi_\theta(W) - F\varphi_\zeta(A,Z)\|^2 + \lambda_1\|F\|^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt $\varphi_\theta, \varphi_\zeta$.

Plug $\hat{F}_{\theta,\zeta}$ into S1 loss, backprop through Cholesky for ζ
(...not θ...why not?)

Xu, Kanagawa, G. (2021).

Final algorithm

Solve for θ, ξ, ζ:

Repeat until convergence:

- Secondary: Solve for $\hat{F}_{\theta,\zeta}$, then gradient steps on ζ (backprop through Cholesky)

Final algorithm

Solve for θ, ξ, ζ:

Repeat until convergence:

- Secondary: Solve for $\hat{F}_{\theta,\zeta}$, then gradient steps on ζ (backprop through Cholesky)
- Primary: Solve for $\hat{\gamma}$ in terms of $\hat{F}_{\theta,\zeta}\varphi_\zeta(A, Z)$ and $\varphi_\xi(A)$

Final algorithm

Solve for θ, ξ, ζ:

Repeat until convergence:

- Secondary: Solve for $\hat{F}_{\theta,\zeta}$, then gradient steps on ζ (backprop through Cholesky)
- Primary: Solve for $\hat{\gamma}$ in terms of $\hat{F}_{\theta,\zeta}\varphi_\zeta(A, Z)$ and $\varphi_\xi(A)$
- Primary: Gradient steps on θ, ξ (backprop through Cholesky)
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.

Final algorithm

Solve for θ, ξ, ζ:

Repeat until convergence:

- Secondary: Solve for $\hat{F}_{\theta,\zeta}$, then gradient steps on ζ (backprop through Cholesky)
- Primary: Solve for $\hat{\gamma}$ in terms of $\hat{F}_{\theta,\zeta}\varphi_\zeta(A, Z)$ and $\varphi_\xi(A)$
- Primary: Gradient steps on θ, ξ (backprop through Cholesky)
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.

Iterate between updates of θ, ξ and ζ

Final algorithm

Solve for θ, ξ, ζ:

Repeat until convergence:

- Secondary: Solve for $\hat{F}_{\theta,\zeta}$, then gradient steps on ζ (backprop through Cholesky)
- Primary: Solve for $\hat{\gamma}$ in terms of $\hat{F}_{\theta,\zeta}\varphi_\zeta(A, Z)$ and $\varphi_\xi(A)$
- Primary: Gradient steps on θ, ξ (backprop through Cholesky)
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.

Iterate between updates of θ, ξ and ζ

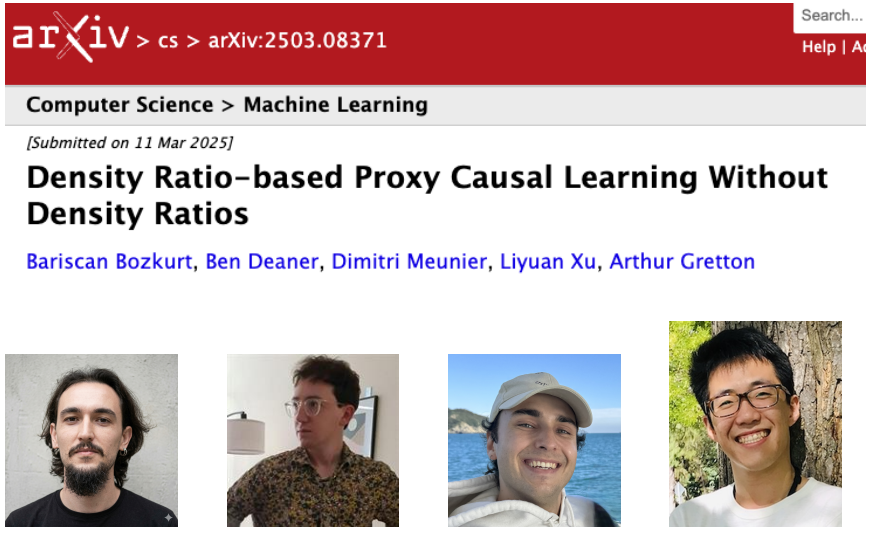
Key point: features $\varphi_\theta(W)$ learned specially for:

$$\mathbb{E}(Y|a, z) = \mathbb{E}_{W|a,z} h_y(W, a)$$

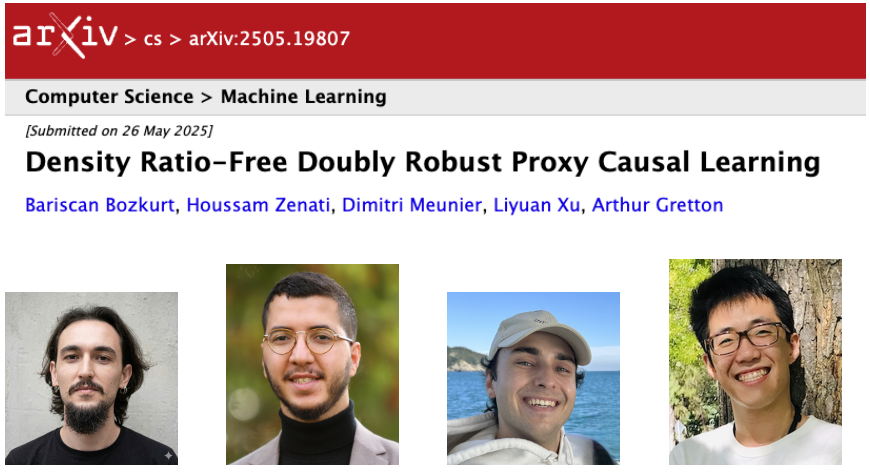
Contrast with autoencoders/sampling: must reconstruct/sample all of W.

Treatment bridge and doubly robust proxy learning

AISTATS 2025



NeurIPS 2025



Code for treatment bridge and doubly robust:

https://github.com/BariscanBozkurt/Doubly-Robust-Kernel-Proxy-Variable-Algorithm

Other DR approaches:

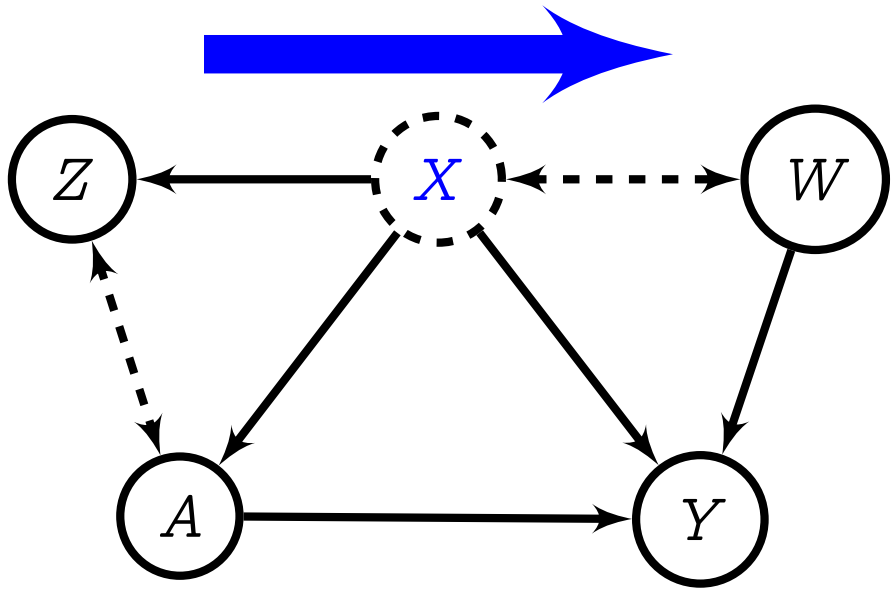
Cui, Pu, Shi, Miao, Tchetgen-Tchetgen. Semiparametric proximal causal inference. JASA (2024): binary treatment.

Wu, Fu, Wang, Sun. Doubly robust proximal causal learning for continuous treatments. ICLR (2024): density ratio + Parzen smoothed treatment

Treatment bridge: idea

Outcome bridge

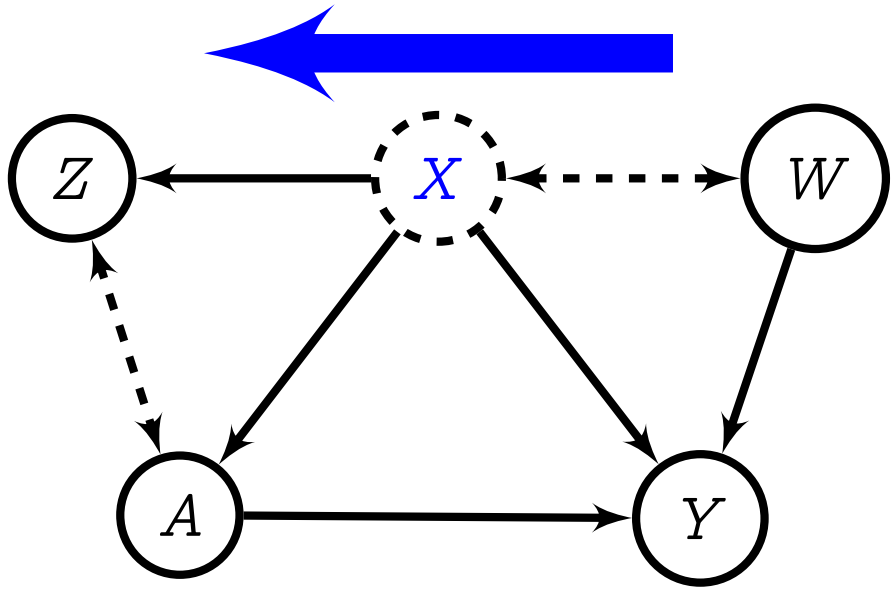
$$\mathbb{E}(Y|a,z) = \mathbb{E}_{W|a,z} h_y(W,a)$$



Treatment bridge: idea

Treatment bridge

$$?? = \mathbb{E}_{Z|a,w}[g_y(Z, a)]$$



Treatment bridge dose-response

Main theorem: Assume we solved for treatment bridge:

$$\frac{p(a)p(w)}{p(a,w)} = \mathbb{E}_{Z|a,w}[g_y(Z,a)]$$

Treatment bridge dose-response

Main theorem: Assume we solved for treatment bridge:

$$\frac{p(a)p(w)}{p(a,w)} = \mathbb{E}_{Z|a,w}[g_y(Z,a)]$$

Dose-response curve via $p(Z|A=a)$:

$$\mathbb{E}(Y^{(a)}) = DR^{(T)}(a; g_y) := \mathbb{E}_{Y,Z|a}[Y\, g_y(Z,a)]$$

Bridge existence requires $\triangle$, identification of DR requires $\diamondsuit$
[BDMXG25: Assumptions 3.2, 3.3, 3.7].

$$W \perp\!\!\!\perp (Z,A)|X$$

$$Y \perp\!\!\!\perp Z|(A,X)$$

$$\mathbb{E}[f(X)|A=a, Z=z] = 0,\ \forall (z,a) \iff f(X) = 0,\ \mathbb{P}_X \text{ a.s.} \quad \triangle$$

$$\mathbb{E}[f(X)|A=a, W=w] = 0,\ \forall (w,a) \iff f(X) = 0,\ \mathbb{P}_X \text{ a.s.} \quad \diamondsuit$$

Orange assumption is slightly stronger than the one we used, and implies our actual Assumption 3.2

Treatment bridge regression loss

Loss function

$$\mathcal{L}(g_y) = \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]$$

Treatment bridge regression loss

Loss function

$$\mathcal{L}(g_y) = \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]$$

$$= \mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] + \text{const.}$$

$$- 2\int \left(\underbrace{\frac{p(a)p(w)}{p(a,w)}}_{\text{density ratio!}}\mathbb{E}_{Z|a,w}[g_y(Z,a)]\right)p(a,w)\,da\,dw$$

Treatment bridge regression loss

Loss function

$$\mathcal{L}(g_y) = \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]$$

$$= \mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] + \text{const.}$$

$$- 2\int \left(\underbrace{\frac{p(a)p(w)}{\cancel{p(a,w)}}}_{\text{density ratio!}}\mathbb{E}_{Z|a,w}[g_y(Z,a)]\right)\cancel{p(a,w)}\,da\,dw$$

Treatment bridge regression loss

Loss function

$$\mathcal{L}(g_y) = \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]$$

$$= \mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] + \text{const.}$$

$$- 2\int \mathbb{E}_{Z|a,w}[g_y(Z,a)]p(a)p(w)\,da\,dw$$

Treatment bridge regression loss

Loss function

$$\mathcal{L}(g_y) = \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]$$

$$= \mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] + \text{const.}$$

$$- 2\int \mathbb{E}_{Z|a,w}[g_y(Z,a)]p(a)p(w)\,da\,dw$$

$$= \underbrace{\mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]}_{(1)} - \underbrace{2\mathbb{E}_A\mathbb{E}_W[\mathbb{E}_{Z|A,W}[g_y(Z,A)]]}_{(2)} + \text{const.}$$

Empirical estimates:

1. Squared mean
2. U-statistic

Feature paramatrization of bridge $g_y(z, a)$

The treatment bridge is a function of two arguments

$$g_y(z, a) = \eta^\top [\varphi_\theta(z) \otimes \varphi_\xi(a)]$$

Feature paramatrization of bridge $g_y(z, a)$

The treatment bridge is a function of two arguments

$$g_y(z, a) = \eta^\top [\varphi_\theta(z) \otimes \varphi_\xi(a)]$$

Thus

$$\mathbb{E}_{Z|a,w}\ g_y(Z, a) = \mathbb{E}_{Z|a,w}\ \left[\eta^\top (\varphi_\theta(Z) \otimes \varphi_\xi(a))\right]$$

Feature paramatrization of bridge $g_y(z, a)$

The treatment bridge is a function of two arguments

$$g_y(z, a) = \eta^\top [\varphi_\theta(z) \otimes \varphi_\xi(a)]$$

Thus

$$\begin{aligned}\mathbb{E}_{Z|a,w}\ g_y(Z, a) = \ \mathbb{E}_{Z|a,w}\ \left[\eta^\top (\varphi_\theta(Z) \otimes \varphi_\xi(a))\right] \\ = \eta^\top \Big(\underbrace{\mathbb{E}_{Z|a,w}\left[\varphi_\theta(Z)\right]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a) \Big)\end{aligned}$$

Feature paramatrization of bridge $g_y(z, a)$

The treatment bridge is a function of two arguments

$$g_y(z, a) = \eta^\top [\varphi_\theta(z) \otimes \varphi_\xi(a)]$$

Thus

$$\begin{aligned}\mathbb{E}_{Z|a,w}\, g_y(Z, a) = \; & \mathbb{E}_{Z|a,w} \left[\eta^\top (\varphi_\theta(Z) \otimes \varphi_\xi(a))\right] \\ & = \eta^\top \Big(\underbrace{\mathbb{E}_{Z|a,w} [\varphi_\theta(Z)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a) \Big)\end{aligned}$$

Need a third regression

$$DR^{(T)}(a; g_y) = \mathbb{E}_{Y,Z|a}[Y\, g_y(Z, a)]$$

Feature paramatrization of bridge $g_y(z, a)$

The treatment bridge is a function of two arguments

$$g_y(z, a) = \eta^\top [\varphi_\theta(z) \otimes \varphi_\xi(a)]$$

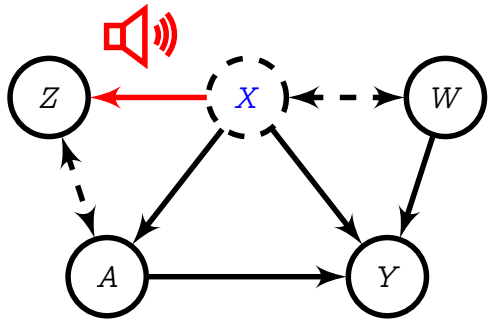
Thus

$$\begin{aligned}\mathbb{E}_{Z|a,w}\ g_y(Z, a) = \ \mathbb{E}_{Z|a,w}\ \left[\eta^\top (\varphi_\theta(Z) \otimes \varphi_\xi(a))\right] \\ = \eta^\top \Big(\underbrace{\mathbb{E}_{Z|a,w}\left[\varphi_\theta(Z)\right]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a)\Big)\end{aligned}$$

Need a third regression

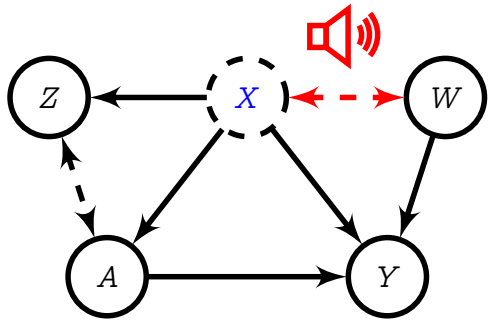
$$DR^{(T)}(a; g_y) = \mathbb{E}_{Y,Z|a}[Y\, g_y(Z, a)] = \eta^\top [\underbrace{\mathbb{E}_{Y,Z|a}[Y\, \varphi_\theta(Z)]}_{\text{cond. feat. mean}} \otimes \varphi_\xi(a)]$$

Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69

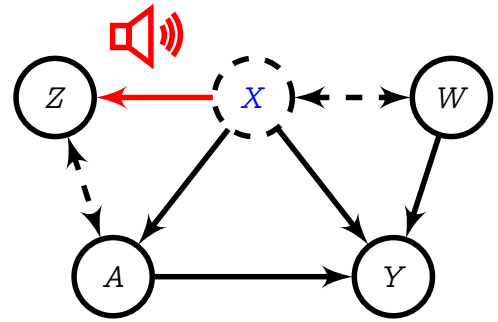
Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

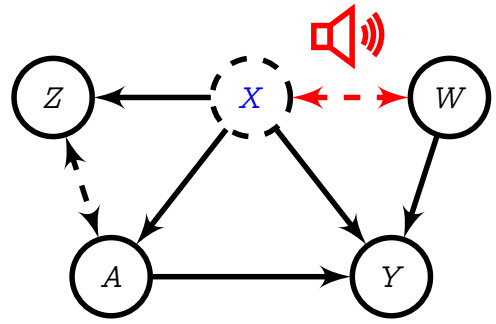
Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29
Set. 3	51.22 ± 46.10	3.47 ± 1.04

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

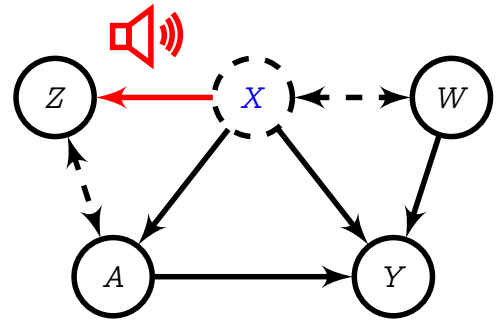
Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29
Set. 3	51.22 ± 46.10	3.47 ± 1.04
Set. 4	8.99 ± 4.91	15.32 ± 5.48

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

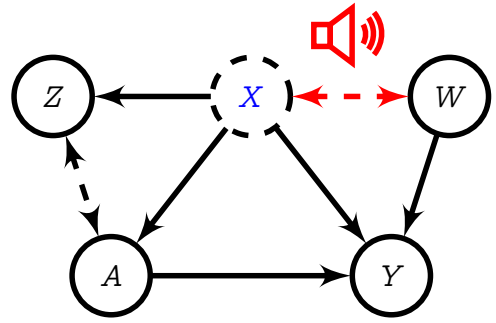
Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29
Set. 3	51.22 ± 46.10	3.47 ± 1.04
Set. 4	8.99 ± 4.91	15.32 ± 5.48
Set. 5	19.82 ± 8.85	11.29 ± 8.23

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

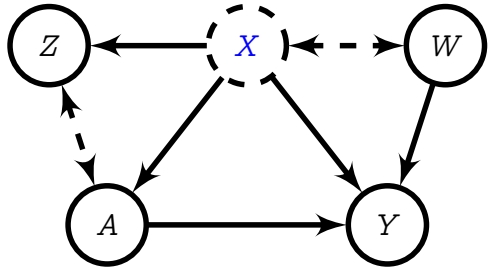
Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29
Set. 3	51.22 ± 46.10	3.47 ± 1.04
Set. 4	8.99 ± 4.91	15.32 ± 5.48
Set. 5	19.82 ± 8.85	11.29 ± 8.23
Set. 6	53.94 ± 15.16	184.36 ± 48.89

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

Why treatment bridge? Synthetic demo



Setting	Outcome bridge	Treatment bridge
Set. 1	41.84 ± 26.61	5.53 ± 0.69
Set. 2	5.41 ± 2.17	9.32 ± 5.29
Set. 3	51.22 ± 46.10	3.47 ± 1.04
Set. 4	8.99 ± 4.91	15.32 ± 5.48
Set. 5	19.82 ± 8.85	11.29 ± 8.23
Set. 6	53.94 ± 15.16	184.36 ± 48.89

Bozkurt, Deaner, Meunier, Xu, G (2025, Section 13.5)
Tsai, Pfohl, Salaudeen, Chiou, Kusner, D'Amour, Koyejo, G. (2024, Appendix D.5)

Doubly robust proxy causal learning

$$DR^{(DR)}(a; h_y, g_y)$$
$$= \mathbb{E}_{Y,Z,W|a}[g_y(Z,a)(Y - h_y(W,a))] + \mathbb{E}_W[h_y(W,a)]$$

Doubly robust proxy causal learning

$$
\begin{aligned}
&DR^{(DR)}(a; h_y, g_y) \\
&= \mathbb{E}_{Y,Z,W|a}[g_y(Z,a)(Y - h_y(W,a))] + \mathbb{E}_W[h_y(W,a)] \\
&= \underbrace{\mathbb{E}_{Y,Z|a}[Y\, g_y(Z,a)]}_{\text{3rd regression}} - \underbrace{\mathbb{E}_{Z,W|a}[g_y(Z,a) h_y(W,a)]}_{\text{4th regression}} + \mathbb{E}_W[h_y(W,a)]
\end{aligned}
$$

Doubly robust proxy causal learning

$$
\begin{aligned}
&DR^{(DR)}(a; h_y, g_y) \\
&= \mathbb{E}_{Y,Z,W|a}[g_y(Z,a)(Y - h_y(W,a))] + \mathbb{E}_W[h_y(W,a)] \\
&= \underbrace{\mathbb{E}_{Y,Z|a}[Y\, g_y(Z,a)]}_{\text{3rd regression}} - \underbrace{\mathbb{E}_{Z,W|a}[g_y(Z,a) h_y(W,a)]}_{\text{4th regression}} + \mathbb{E}_W[h_y(W,a)]
\end{aligned}
$$

Guarantees?

Doubly robust proxy causal learning

$$
\begin{aligned}
&DR^{(DR)}(a; h_y, g_y) \\
&= \mathbb{E}_{Y,Z,W|a}[g_y(Z,a)(Y - h_y(W,a))] + \mathbb{E}_W[h_y(W,a)] \\
&= \underbrace{\mathbb{E}_{Y,Z|a}[Y\, g_y(Z,a)]}_{\text{3rd regression}} - \underbrace{\mathbb{E}_{Z,W|a}[g_y(Z,a) h_y(W,a)]}_{\text{4th regression}} + \mathbb{E}_W[h_y(W,a)]
\end{aligned}
$$

Guarantees?

- If either of h_y, g_y is correct then $DR^{(DR)}(a; h_y, g_y) = DR(a)$ (true dose-response)

Doubly robust proxy causal learning

$$
\begin{aligned}
&DR^{(DR)}(a; h_y, g_y) \\
&= \mathbb{E}_{Y,Z,W|a}[g_y(Z,a)(Y - h_y(W,a))] + \mathbb{E}_W[h_y(W,a)] \\
&= \underbrace{\mathbb{E}_{Y,Z|a}[Y\, g_y(Z,a)]}_{\text{3rd regression}} - \underbrace{\mathbb{E}_{Z,W|a}[g_y(Z,a) h_y(W,a)]}_{\text{4th regression}} + \mathbb{E}_W[h_y(W,a)]
\end{aligned}
$$

Guarantees?

- If either of h_y, g_y is correct then $DR^{(DR)}(a; h_y, g_y) = DR(a)$ (true dose-response)
- Convergence:

$$
\begin{aligned}
&|DR(a) - \widehat{DR}^{(DR)}(a; \hat{h}_y, \hat{g}_y)| \\
&\lesssim \|g_y - \hat{g}_y\| \|h_y - \hat{h}_y\| + \|C_{YZ|A} - \widehat{C}_{YZ|A}\|_{HS} + \|C_{WZ|A} - \widehat{C}_{WZ|A}\|_{HS}
\end{aligned}
$$

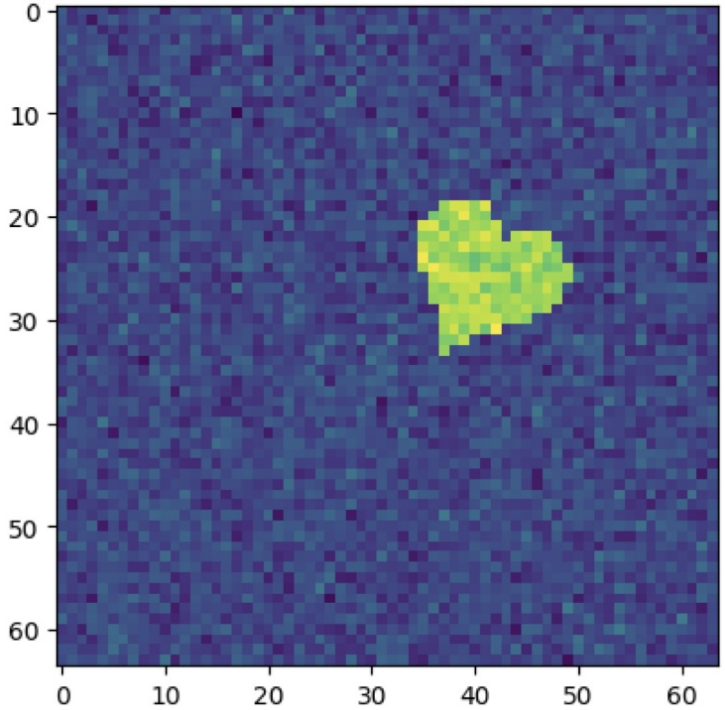
where

$$
C_{YZ|A}\varphi(a) = \mathbb{E}_{Y,Z|a}[Y\varphi(Z)] \qquad C_{WZ|A}\varphi(a) = \mathbb{E}_{W,Z|a}[\varphi(W) \otimes \varphi(Z)]
$$

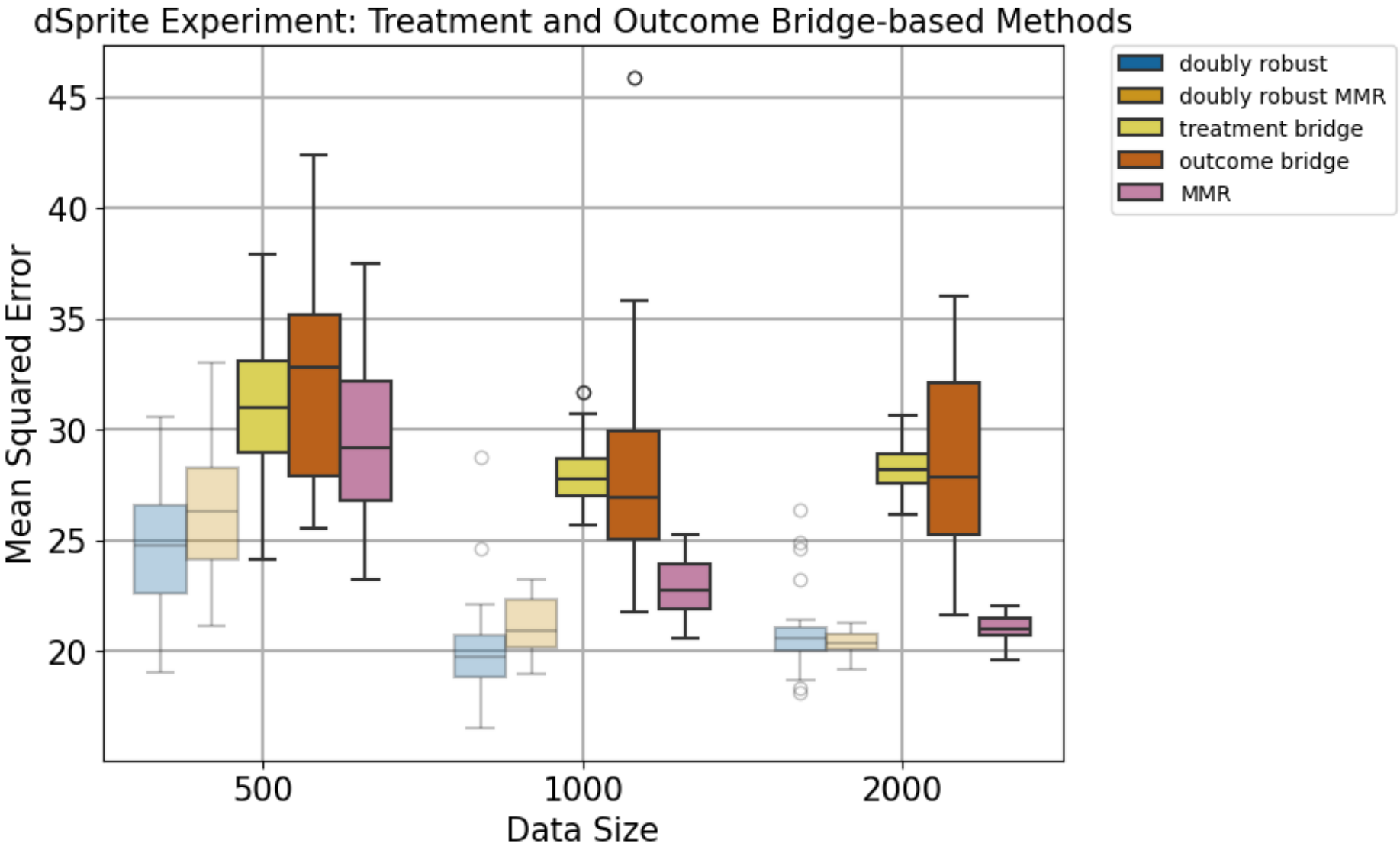
Synthetic experiment, kernel features

dSprite example:

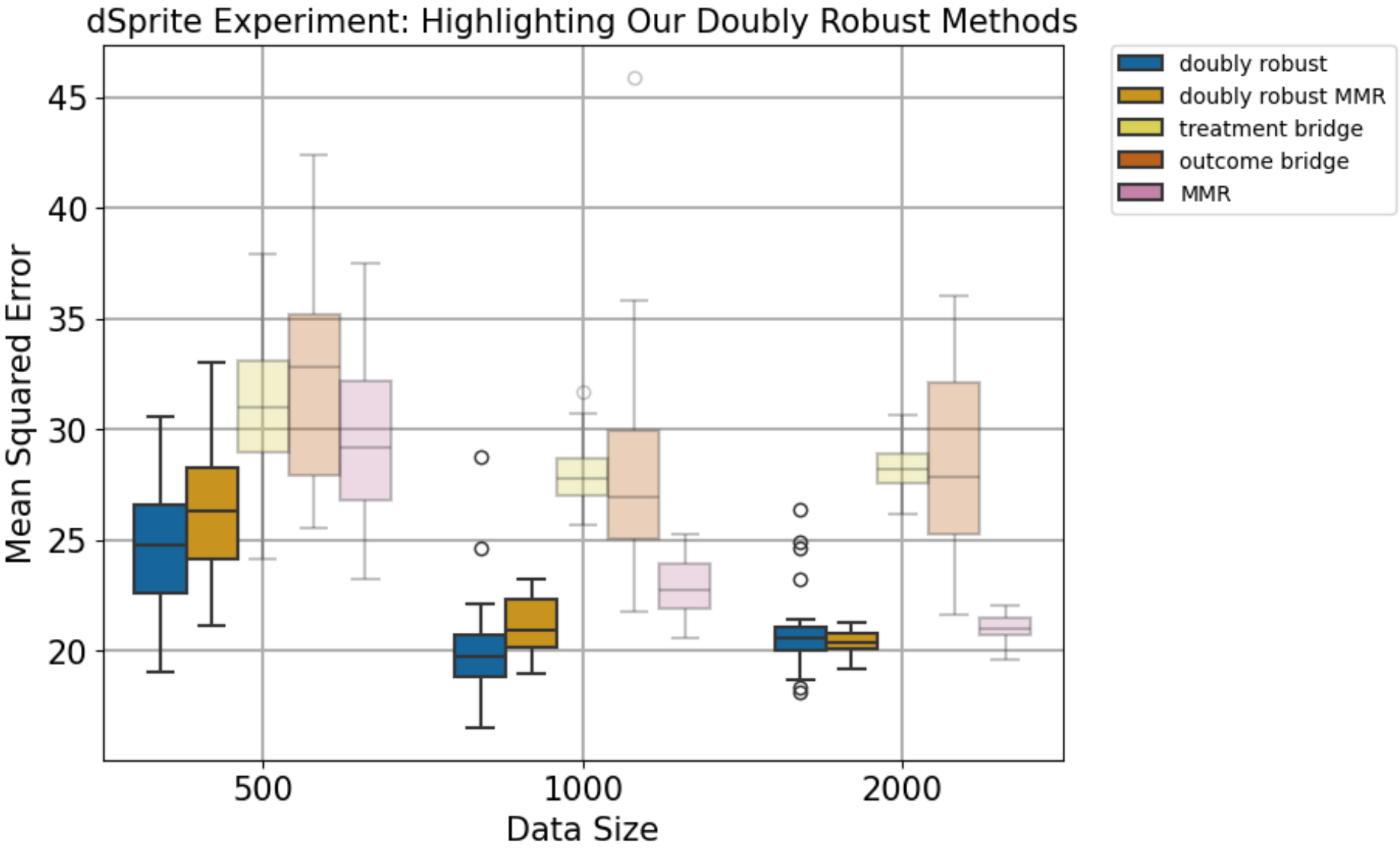
- $X = \{\texttt{scale}, \texttt{rotation}, \texttt{posX}, \texttt{posY}\}$
- Treatment $A \in \mathbb{R}^{4096}$ is the image generated (with Gaussian noise)
- Outcome Y is linear function of A with multiplicative confounding by $\texttt{posY}$.
- $Z = \{\texttt{scale}, \texttt{rotation}, \texttt{posX}\}$, $W =$ noisy image sharing $\texttt{posY}$



Synthetic experiment, results



Synthetic experiment, results

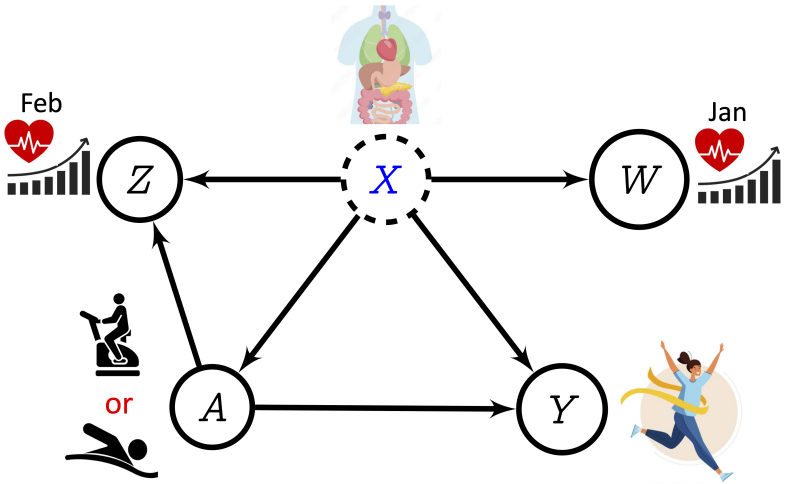


Conclusion

Causal effect estimation with unobserved X, (possibly) complex nonlinear effects on A, Y

We need to observe:

- Treatment proxy Z (interacts with A, but not directly with Y)
- Outcome proxy W (no direct interaction with A, can affect Y)

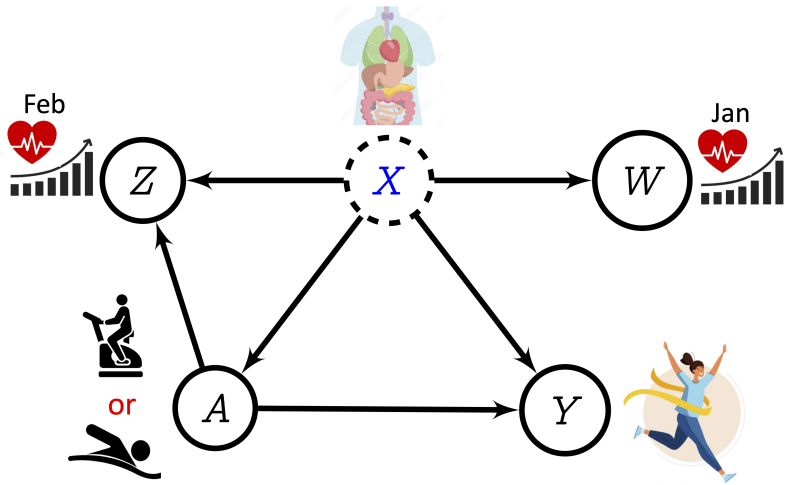


Conclusion

Causal effect estimation with unobserved X, (possibly) complex nonlinear effects on A, Y

We need to observe:

- Treatment proxy Z (interacts with A, but not directly with Y)
- Outcome proxy W (no direct interaction with A, can affect Y)



Key messages:

- Don't ~~meet your heroes~~ model/sample latents X
- Don't model all of W, only relevant features for Y
- "Ridge regression is all you need"

Code available:

https://github.com/liyuan9988/DeepFeatureProxyVariable/

https://github.com/BariscanBozkurt/Doubly-Robust-Kernel-Proxy-Variable-Algorithm

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Google DeepMind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT