Causal Effect Estimation using Proxy Variables

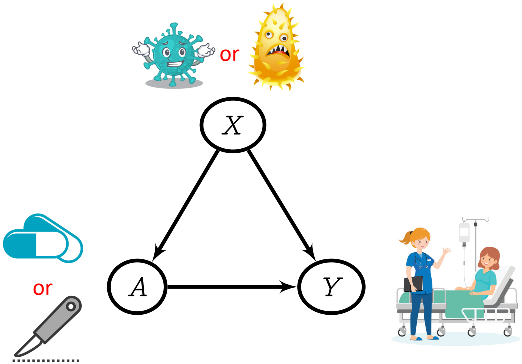
Arthur Gretton

Gatsby Computational Neuroscience Unit
Google DeepMind

LMS Research Summer School, University of Warwick

Observation vs intervention

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

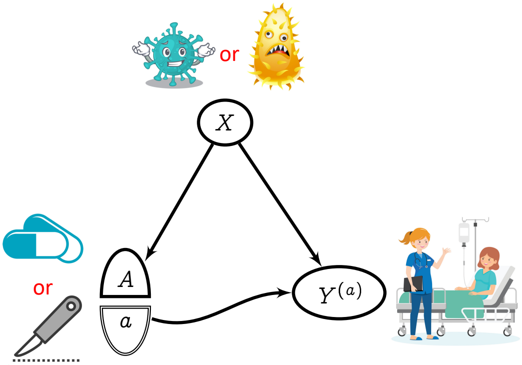


From our *observations* of historical hospital data:

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

Observation vs intervention

Average causal effect (intervention): $\mathbb{E}[Y^{(a)}] = \sum_x \mathbb{E}[Y \mid a, x]p(x)$

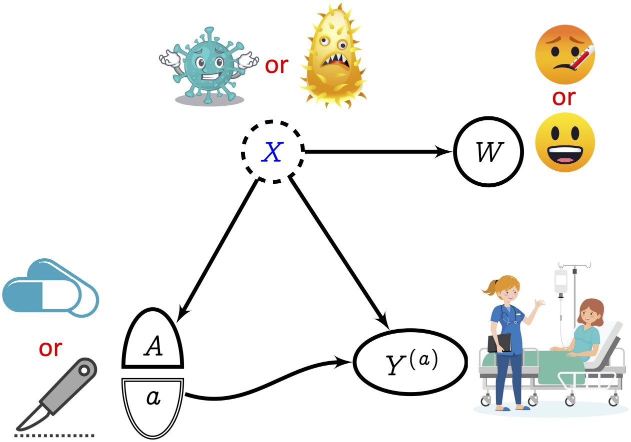


From our *intervention* (making all patients take a treatment):

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

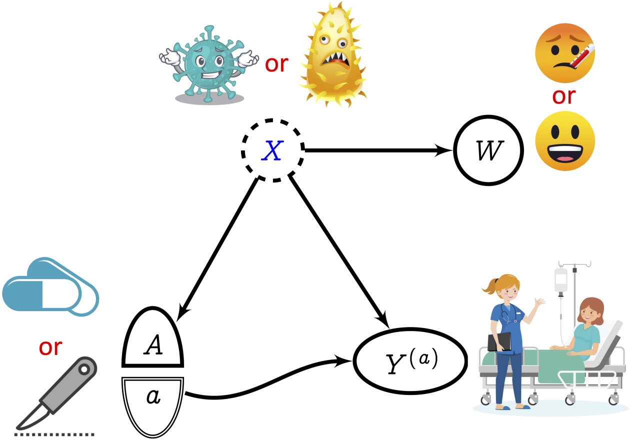
Richardson, Robins (2013), Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

We record symptom W, not disease X



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

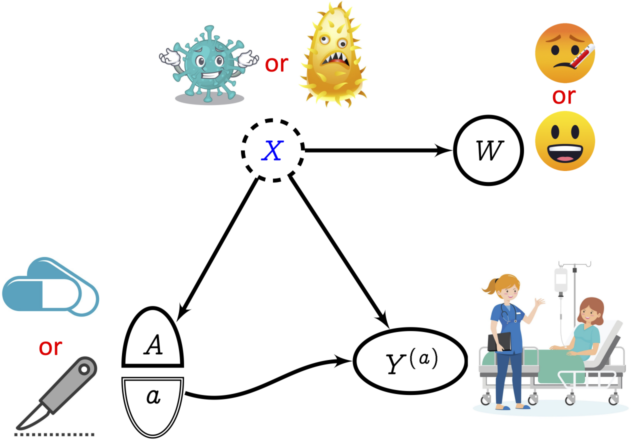
We record symptom W, not disease X



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

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

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

Pearl (2010), On Measurement Bias in Causal Inference

Outline

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)

Related topic: domain shift (the blessings of multiple domains)

Outline

Causal effect estimation, with hidden covariates X:

- Use proxy variables (negative controls)

Applications: effect of actions under

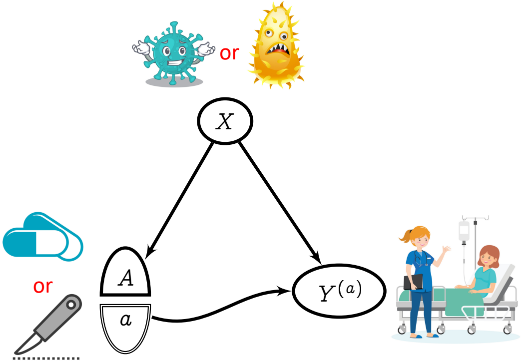
- privacy constraints (email, ads, DMA)
- data gathering constraints (edge computing)
- fundamental limitations (preferences, state of mind)

Related topic: domain shift (the blessings of multiple domains)

What's new and why?

- Treatment A, proxy variables, etc can be multivariate, complicated...
- ...by using kernel or neural net feature representations
- Don't ~~meet your heroes~~ model your hidden variables!

Some core assumptions



Assume:

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

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.

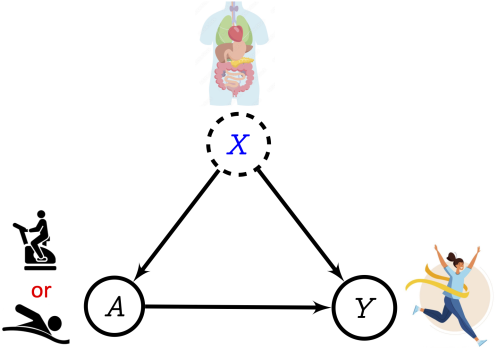
etchetge@hsph.harvard.edu

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

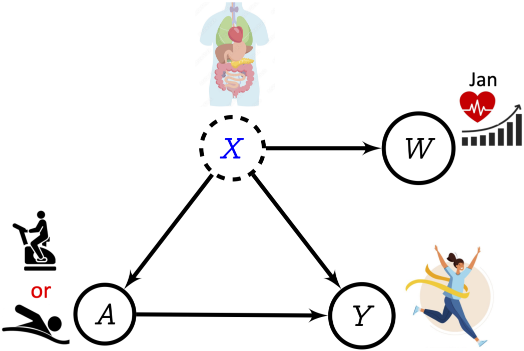


What are proxies, and when are they useful?

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

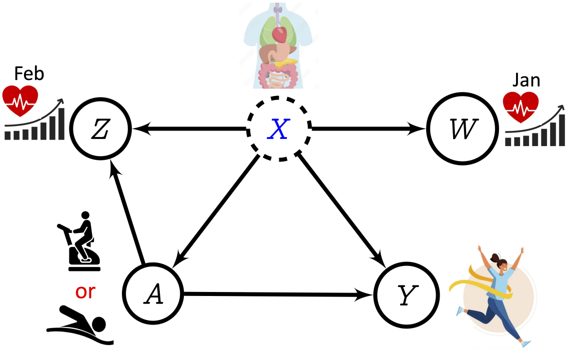


What are proxies, and when are they useful?

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

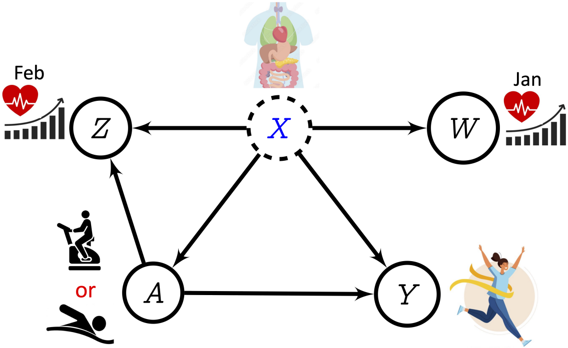


What are proxies, and when are they useful?

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



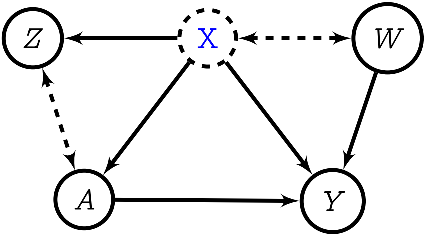
$\Longrightarrow$ Can recover $\mathbb{E}(Y^{(a)})$ from observational data

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



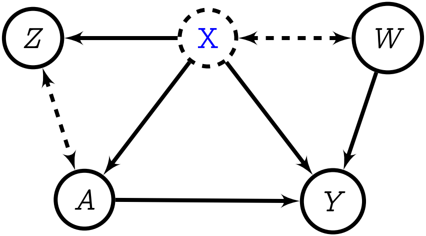
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

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:

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

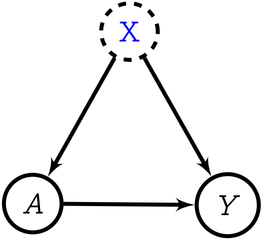
$$Y \perp\!\!\!\perp Z | (A, X)$$

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



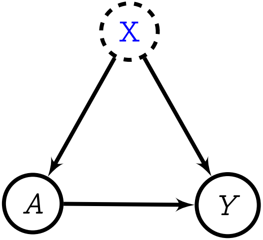
If X were observed,

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

Why proxy variables? A simple proof

The definitions are:

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



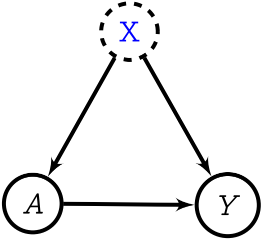
If X were observed,

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

Why proxy variables? A simple proof

The definitions are:

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



If X were observed,

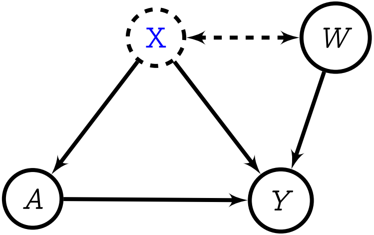
$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

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



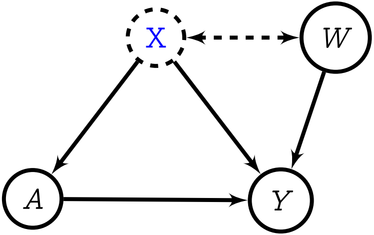
For each a, if we could solve:

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

...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:

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

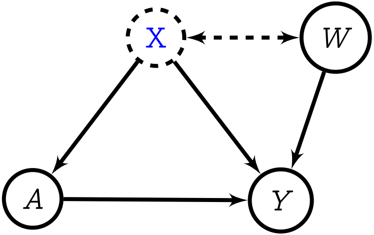
.....then

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

...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:

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

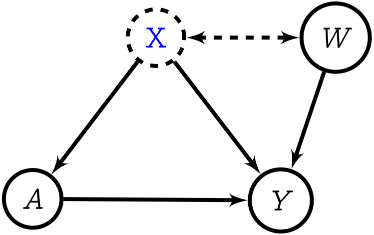
.....then

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

...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:

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

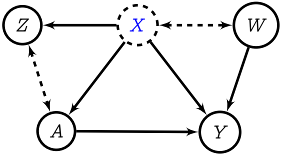
.....then

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

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

From last slide,

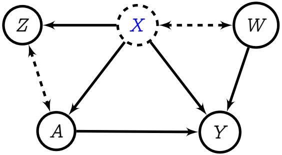
$$P(Y|X,a) = H_{w,a} P(W|X)$$



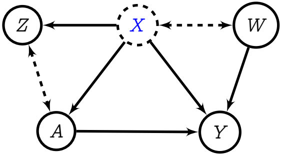
...now project onto $p(X|Z,a)$

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 $p(X|Z,a)$



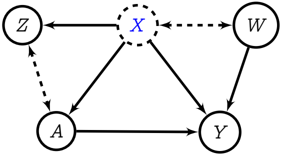
From last slide,

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

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,

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

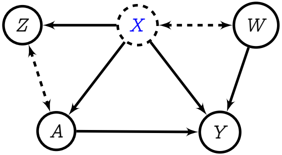
Because $W \perp\!\!\!\perp (Z,A)|X$,

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

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

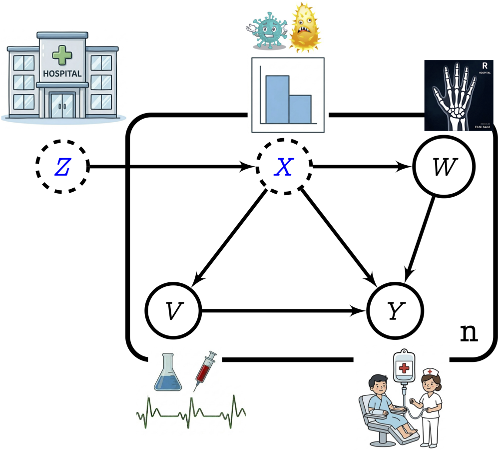
"The blessings of multiple domains"

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.)
- $n \in \{1 \ldots D+1\}$




arXiv > cs > arXiv:2403.07442

Computer Science > Machine Learning

[Submitted on 12 Mar 2024]

Proxy Methods for Domain Adaptation

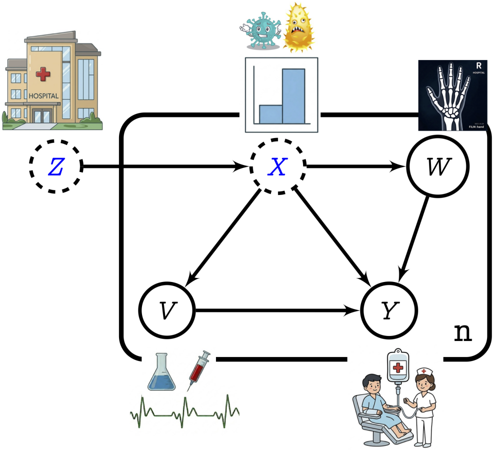
Katherine Tsai, Stephen R. Pfohl, Olawale Salaudeen, Nicole Chiou, Matt J. Kusner, Alexander D'Amour, Sanmi Koyejo, Arthur Gretton

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.)
- $n \in \{1 \ldots D+1\}$




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

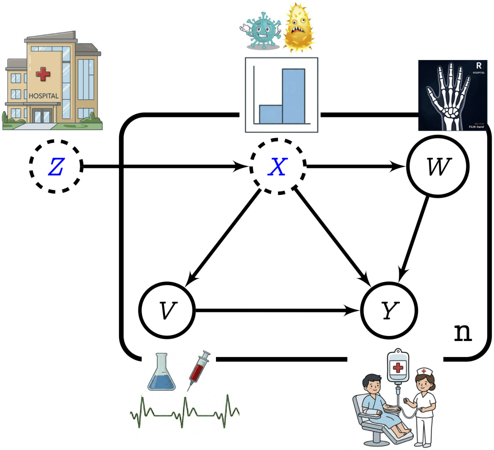
Outcome bridge for domain shift

"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.)
- $n \in \{1 \ldots D+1\}$

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



arXiv > cs > arXiv:2403.07442

Computer Science > Machine Learning

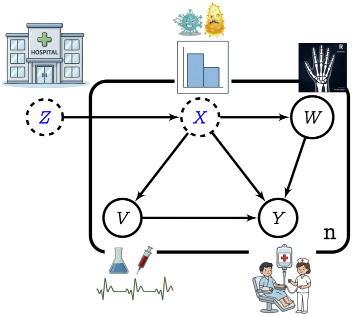
[Submitted on 12 Mar 2024]

Proxy Methods for Domain Adaptation

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

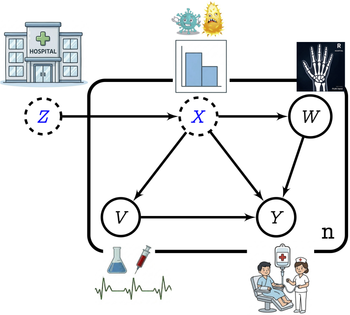
Proof (1)

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



Proof (1)

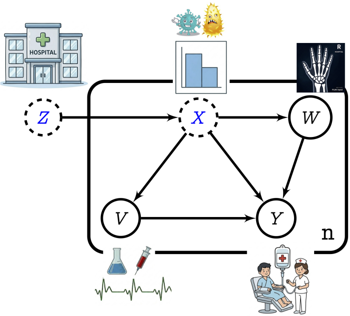
- X: unobserved variable.
- V: "view 1"
- W: "view 2"
- Y: outcome
- n observations per domain



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



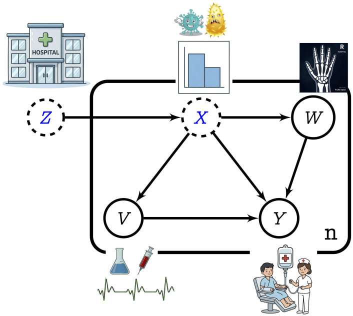
Core assumption: only $P(X|Z)$ changes with Z, all $P(Q|X)$ invariant for $Q \subseteq \{V, Y, W\}$.

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

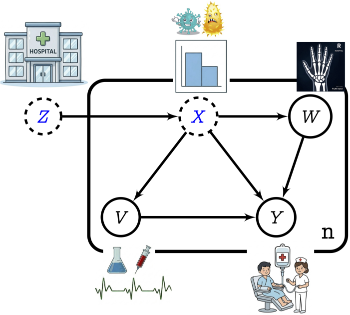
$$P(Y|X, v) = M_{w,v} P(W|X)$$

....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

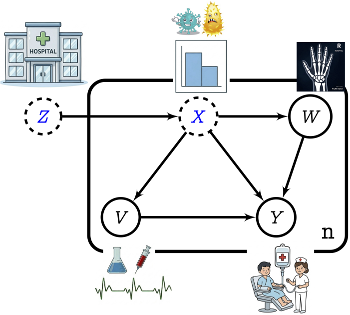
$$P(Y|X, v) = M_{w,v} P(W|X)$$

....then

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

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

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

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

Get rid of the blue!

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

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \begin{bmatrix} P^{(1)}(X|v) & \dots & P^{(D)}(X|v) \end{bmatrix}$$

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

Get rid of the blue!

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

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \begin{bmatrix} P^{(1)}(X|v) & \dots & P^{(D)}(X|v) \end{bmatrix}$$

Then

$$\begin{aligned} P^{(1..D)}(W|v) &:= \begin{bmatrix} P^{(1)}(W|v) & \dots & P^{(D)}(W|v) \end{bmatrix} \\ &= P(W|X)P^{(1..D)}(X|v) \end{aligned}$$

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

Get rid of the blue!

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

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \begin{bmatrix} P^{(1)}(X|v) & \dots & P^{(D)}(X|v) \end{bmatrix}$$

Then

$$\begin{aligned} P^{(1..D)}(W|v) &:= \begin{bmatrix} P^{(1)}(W|v) & \dots & P^{(D)}(W|v) \end{bmatrix} \\ &= P(W|X)P^{(1..D)}(X|v) \end{aligned}$$

Similarly

$$\begin{aligned} P^{(1..D)}(Y|v) &:= \begin{bmatrix} P^{(1)}(Y|v) & \dots & P^{(D)}(Y|v) \end{bmatrix} \\ &= P(Y|X,v)P^{(1..D)}(X|v) \end{aligned}$$

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

Get rid of the blue!

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

$$\underbrace{P^{(1..D)}(X|v)}_{d_x \times D} := \begin{bmatrix} P^{(1)}(X|v) & \dots & P^{(D)}(X|v) \end{bmatrix}$$

Then

$$\begin{aligned} P^{(1..D)}(W|v) &:= \begin{bmatrix} P^{(1)}(W|v) & \dots & P^{(D)}(W|v) \end{bmatrix} \\ &= P(W|X)P^{(1..D)}(X|v) \end{aligned}$$

Similarly

$$\begin{aligned} P^{(1..D)}(Y|v) &:= \begin{bmatrix} P^{(1)}(Y|v) & \dots & P^{(D)}(Y|v) \end{bmatrix} \\ &= P(Y|X,v)P^{(1..D)}(X|v) \end{aligned}$$

Then

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

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

Proxy/Negative Control Methods in the Real World

Outcome bridge and proxy variables

Kernel features (ICML 2021):


arXiv.org > cs > arXiv:2105.04544

Search...

Help | Advan

Computer Science > Machine Learning

[Submitted on 10 May 2021 (v1), last revised 9 Oct 2021 (this version, v4)]

Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, Krikamol Muandet







NN features (NeurIPS 2021):


arXiv.org > cs > arXiv:2106.03907

Search...

Help | Advanc

Computer Science > Machine Learning

[Submitted on 7 Jun 2021 (v1), last revised 7 Dec 2021 (this version, v2)]

Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Liyuan Xu, Heishiro Kanagawa, Arthur Gretton


Code for NN and kernel proxy methods:

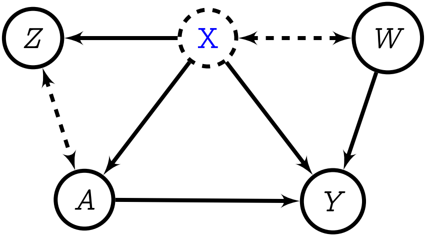
https://github.com/liyuan9988/DeepFeatureProxyVariable/

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, outcome bridge

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, outcome bridge

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 outcome bridge:

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

- "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. Bridge existence requires $\diamondsuit$, identification of DR 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, outcome bridge

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 outcome bridge:

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

- "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 $p(w)$:

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

Proxy relation, outcome bridge

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 outcome bridge:

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

- "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 $p(w)$:

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

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:

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

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:

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

Why?

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

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

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:

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

Why?

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

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

...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).

Feature paramatrization of bridge $h_y(a, w)$

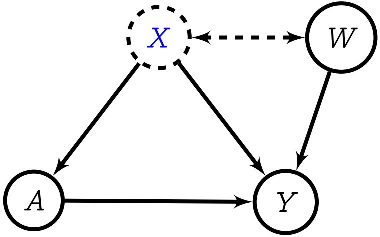
The outcome bridge function class $\mathcal{H}$ defined as:

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

Assume we have:

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

(argument of feature map indicates feature space)



Feature paramatrization of bridge $h_y(a, w)$

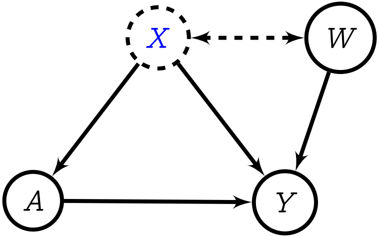
The outcome bridge function class $\mathcal{H}$ defined as:

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

Assume we have:

- output proxy kernel/NN features $\varphi_\theta(w)$
- treatment kernel/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)!

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda \Omega(\|\gamma\|^2)$$

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

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda \Omega(\|\gamma\|^2)$$

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

Density estimation for $p(W|a, z)$? Sample from $p(W|a, z)$?

Deaner (2021).

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

Xu, Kanagawa, G. (2021).

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h\in\mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda\Omega(\|\gamma\|^2)$$

Recall bridge function

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

Deaner (2021).

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

Xu, Kanagawa, G. (2021).

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda \Omega(\|\gamma\|^2)$$

Recall bridge function

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

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

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda \Omega(\|\gamma\|^2)$$

Recall bridge function

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

(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).

Regression for $h_y(w, a)$

Goal:

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

Primary regression:

$$h_y^{(\lambda)} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{Y,A,Z} \left(Y - \mathbb{E}_{W|A,Z} h(W, A) \right)^2 + \lambda \Omega(\|\gamma\|^2)$$

Recall bridge function

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

Ridge regression (again!): 2SLS

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

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

Treatment bridge and doubly robust proxy learning

AISTATS 2025

arXiv > cs > arXiv:2503.08371

Search...
Help | Ac

Computer Science > Machine Learning

[Submitted on 11 Mar 2025]

Density Ratio-based Proxy Causal Learning Without Density Ratios

Bariscan Bozkurt, Ben Deaner, Dimitri Meunier, Liyuan Xu, Arthur Gretton




NeurIPS 2025

arXiv > cs > arXiv:2505.19807

Computer Science > Machine Learning

[Submitted on 26 May 2025]

Density Ratio-Free Doubly Robust Proxy Causal Learning

Bariscan Bozkurt, Houssam Zenati, Dimitri Meunier, Liyuan Xu, Arthur Gretton




Code for treatment bridge and doubly robust:

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

Other DR approaches:

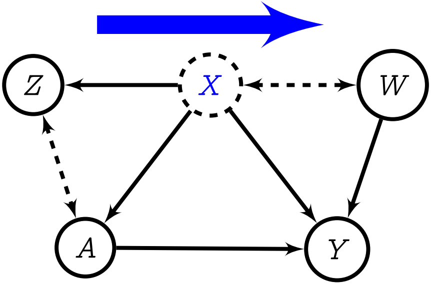
Cui, Pu, Shi, Miao, Tchetgen-Tchetgen. Semiparametric proximal causal inference. JASA (2024): binary treatment.

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

Treatment bridge: idea

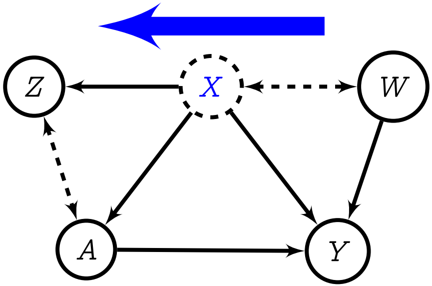
Outcome bridge

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



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].

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

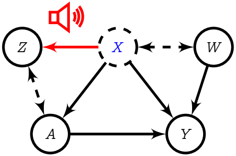
$$Y \perp\!\!\!\perp Z|(A,X)$$

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

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

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

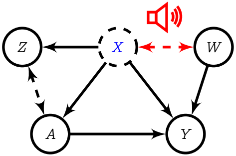
Why treatment bridge? Synthetic demo



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

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

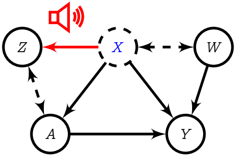
Why treatment bridge? Synthetic demo



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

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

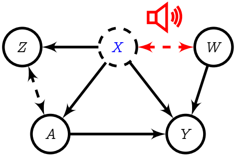
Why treatment bridge? Synthetic demo



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)

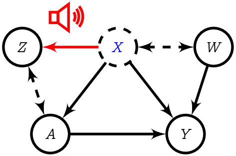
Why treatment bridge? Synthetic demo



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)

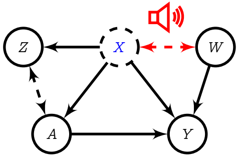
Why treatment bridge? Synthetic demo



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)

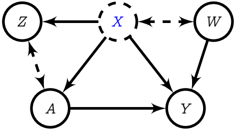
Why treatment bridge? Synthetic demo



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

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

Doubly robust proxy causal learning

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

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:

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

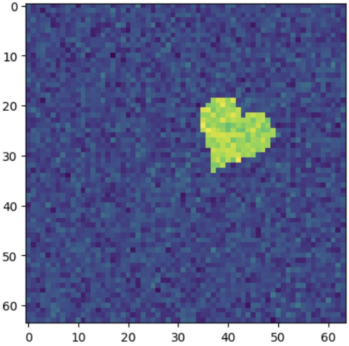
where

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

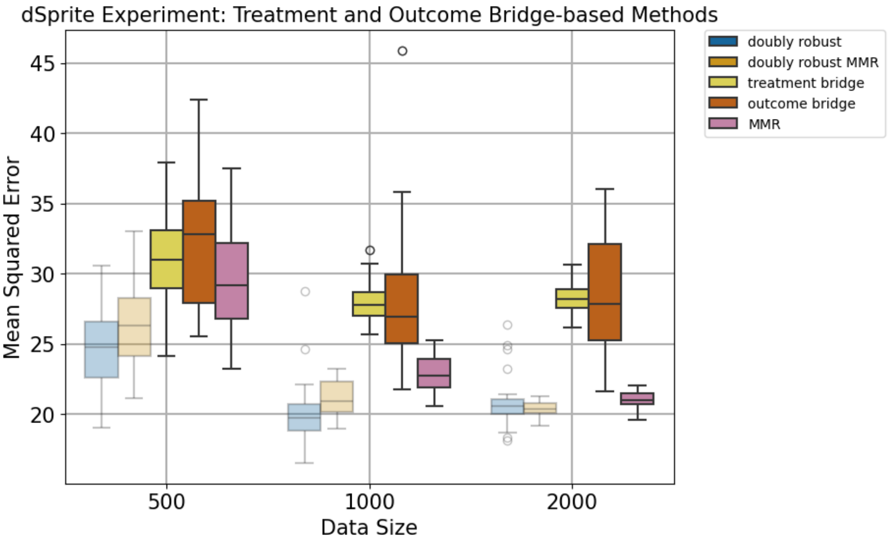
Synthetic experiment, kernel features

dSprite example:

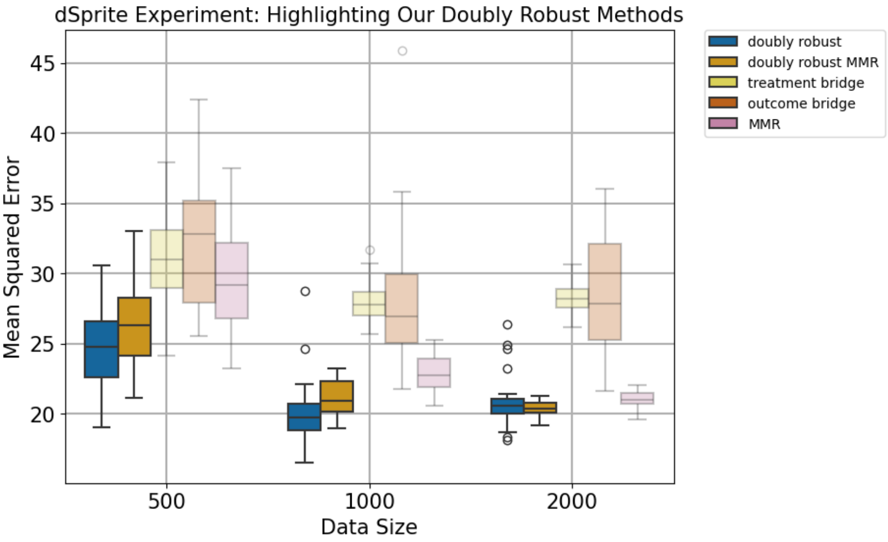
- $X = \{\mathtt{scale}, \mathtt{rotation}, \mathtt{posX}, \mathtt{posY}\}$
- Treatment $A \in \mathbb{R}^{4096}$ is the image generated (with Gaussian noise)
- Outcome Y is linear function of A with multiplicative confounding by posY.
- $Z = \{\mathtt{scale}, \mathtt{rotation}, \mathtt{posX}\}$,
  W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



Synthetic experiment, results



Synthetic experiment, results

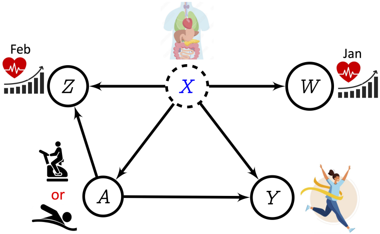


Conclusion

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)

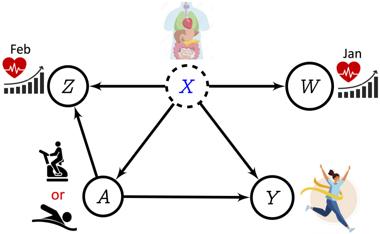


Conclusion

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

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

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

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

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

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

Treatment bridge regression loss

Loss function

$$\begin{aligned}\mathcal{L}(g_y) &= \mathbb{E}_{A,W}\left[\left(\frac{p(A)p(W)}{p(A,W)} - \mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] \\ &= \mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right] + \text{const.} \\ &\qquad - 2\int \mathbb{E}_{Z|a,w}[g_y(Z,a)]p(a)p(w)dadw \\ &= \underbrace{\mathbb{E}_{A,W}\left[\left(\mathbb{E}_{Z|A,W}[g_y(Z,A)]\right)^2\right]}_{(1)} - \underbrace{2\mathbb{E}_A\mathbb{E}_W[\mathbb{E}_{Z|A,W}[g_y(Z,A)]]}_{(2)} + \text{const.}\end{aligned}$$

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

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

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

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

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

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

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

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

Need a third regression

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

Failures of completeness assumptions (1)

Recall (one of the) completeness assumptions:

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

For conciseness, assume conditioning on some a.

Failure 1: $Z \perp\!\!\!\perp X$ (no information about X in proxy)

$$g(X|) = \tilde{g}(X) - \mathbb{E}_X \tilde{g}(X)$$
$$\mathbb{E}(g(X)|Z,a) = \mathbb{E}g(X) = 0.$$

Failures of identifiability assumptions (2)

Failure 2: "exploitable invariance" of $p(X|z)$

$$X \sim \mathcal{N}(0,1),$$
$$Z = |X| + \mathcal{N}(0,1),$$

where $p(X|z) \propto p(z|X)p(X)$ symmetric in X. Consider square integrable *antisymmetric* function $g(X) = -g(-X) \neq 0$. Then

$$\begin{aligned}
\mathbb{E}[g(X)|Z=z] &= \int_{-\infty}^{\infty} g(X)p(X|z)dX \\
&= \int_{-\infty}^{0} g(X)p(X|z)dX + \int_{0}^{\infty} g(X)p(X|z)dX \\
&= 0.
\end{aligned}$$

If distribution of $X|Z$ retains the same "symmetry class" over a set of Z with nonzero measure, then the assumption is violated by $g(X)$ with zero mean on this class.