Causal Effect Estimation with Context and Confounders

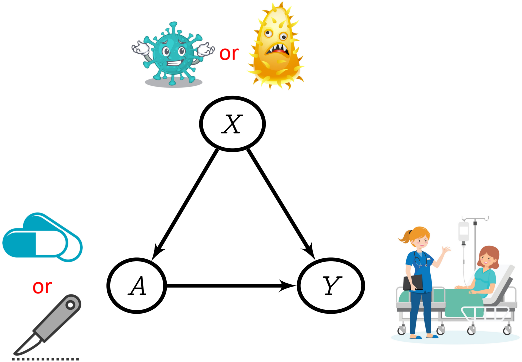
Arthur Gretton

Gatsby Computational Neuroscience Unit
Google Deepmind

Advanced Topics in Machine Learning, 2023

Observation vs intervention

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

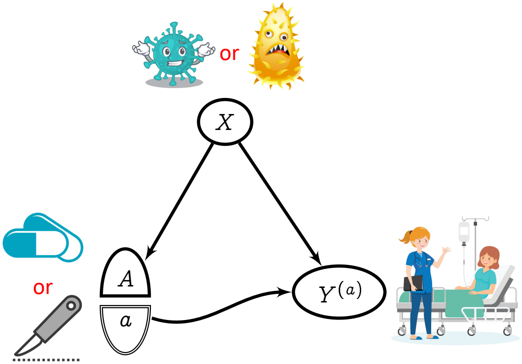


From our *observations* of historical hospital data:

- $P(Y = \text{cured}|A = \text{pills}) = 0.80$
- $P(Y = \text{cured}|A = \text{surgery}) = 0.72$

Observation vs intervention

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

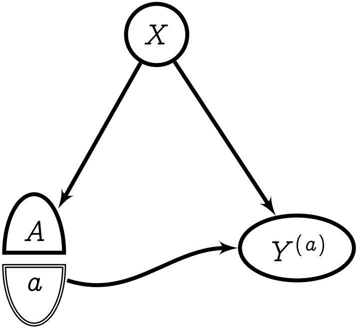


From our *intervention* (making all patients take a treatment):

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

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

Questions we will solve



Outline

Causal effect estimation, observed covariates:

- Average treatment effect (ATE), *conditional* average treatment effect (CATE)

Causal effect estimation, hidden covariates:

- ... proxy variables

What's new? What is it good for?

- Treatment A, covariates X, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations

One model: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x) = \langle \gamma, \varphi_\theta(x) \rangle_{\mathcal{H}}$$

One model: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x) = \langle \gamma, \varphi_\theta(x) \rangle_{\mathcal{H}}$$

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

Xu, G., A Neural mean embedding approach for back-door and front-door adjustment. (ICLR 23)

Xu, Chen, Srinivasan, de Freitas, Doucet, G. Learning Deep Features in Instrumental Variable Regression. (ICLR 21)

Xu, Kanagawa, G. "Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation". (NeurIPS 21)

Kernel approach: Infinite dictionaries of fixed kernel features:

$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Kernel is feature dot product.

Singh, Xu, G. Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. (Biometrika, 2023)

Singh, Sahani, G. Kernel Instrumental Variable Regression. (NeurIPS 19)

Mastouri*, Zhu*, Gultchin, Korba, Silva, Kusner, G,[†] Muandet[†] (2021); Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction (ICML21)

Model fitting: *kernel* ridge regression

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

$$\hat{\gamma} \quad = \quad \arg\min_{\gamma \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle \gamma, \varphi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|_{\mathcal{H}}^2 \right).$$

Kernel solution at x

(as weighted sum of y)

$$\hat{\gamma}(x) = \sum_{i=1}^{n} y_i \beta_i(x)$$

$$\beta(x) = (K_{XX} + \lambda I)^{-1} k_{Xx}$$

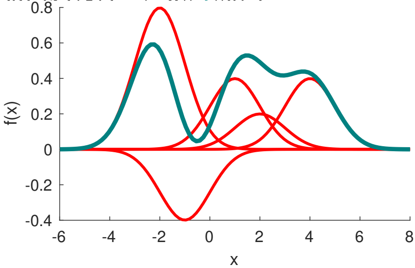
$$(K_{XX})_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$$

$$(k_{Xx})_i = k(x_i, x)$$

Model fitting: *kernel* ridge regression

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

$$\hat{\gamma} = \arg\min_{\gamma\in\mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle \gamma, \varphi(x_i)\rangle_{\mathcal{H}})^2 + \lambda\|\gamma\|^2_{\mathcal{H}} \right)$$



Kernel solution at x
(as weighted sum of y)

$$\hat{\gamma}(x) = \sum_{i=1}^{n} y_i \beta_i(x)$$

$$\beta(x) = (K_{XX} + \lambda I)^{-1} k_{Xx}$$

$$(K_{XX})_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j)\rangle_{\mathcal{H}}$$

$$(k_{Xx})_i = k(x_i, x)$$

Observed covariates: (conditional) ATE

Kernel (Biometrika 2023):


arXiv > econ > arXiv:2010.04855

Search...

Help | Advan

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton

NN (ICLR 2023):


arXiv > cs > arXiv:2210.06610

Search...

Help | Advan

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/

Observed covariates: (conditional) ATE

Kernel features (in revision, Biometrika):

arXiv > econ > arXiv:2010.04855

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 23 Aug 2022 (this version, v6)]

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

Rahul Singh, Liyuan Xu, Arthur Gretton

NN features (ICLR 2023):

arXiv > cs > arXiv:2210.06610

Computer Science > Machine Learning

[Submitted on 12 Oct 2022]

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment

Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/

Average treatment effect

Potential outcome (intervention):

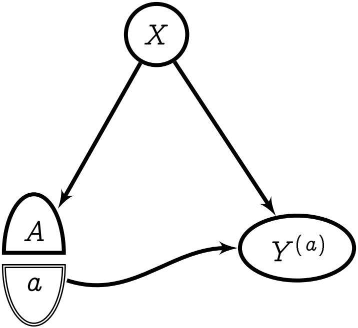
$$\mathbb{E}[Y^{(a)}] = \int \mathbb{E}[Y|a,x]\,dp(x)$$

(the average structural function; in epidemiology, for continuous a, the dose-response curve).

Assume: (1) Stable Unit Treatment Value Assumption (aka “no interference”), (2) Conditional exchangeability $Y^{(a)} \perp\!\!\!\perp A|X$. (3) Overlap.

Example: US job corps, training for disadvantaged youths:

- A: treatment (training hours)
- Y: outcome (percentage employment)
- X: covariates (age, education, marital status, ...)



Multiple inputs via products of kernels

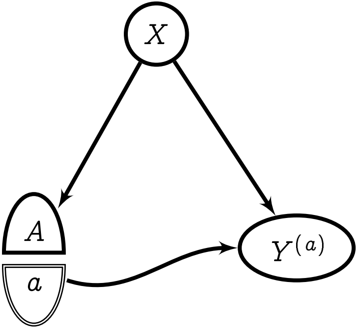
We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[Y | a, x]$$

Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

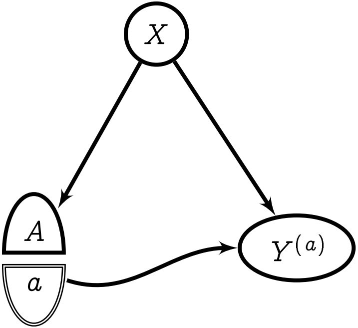
(argument of kernel/feature map indicates feature space)



Multiple inputs via products of kernels

We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[Y | a, x]$$



Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

(argument of kernel/feature map indicates feature space)

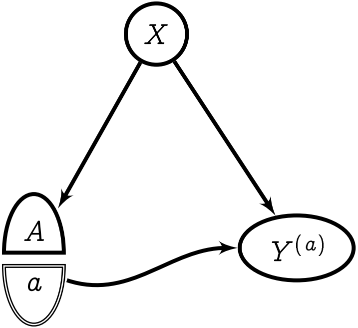
We use outer product of features ($\implies$ product of kernels):

$$\phi(x, a) = \varphi(a) \otimes \varphi(x) \qquad \mathfrak{K}([a, x], [a', x']) = k(a, a')k(x, x')$$

Multiple inputs via products of kernels

We may predict expected outcome from two inputs

$$\gamma_0(a, x) := \mathbb{E}[\,Y\,|\,a, x\,]$$



Assume we have:

- covariate features $\varphi(x)$ with kernel $k(x, x')$
- treatment features $\varphi(a)$ with kernel $k(a, a')$

(argument of kernel/feature map indicates feature space)

We use outer product of features ($\implies$ product of kernels):

$$\phi(x, a) = \varphi(a) \otimes \varphi(x) \qquad \mathfrak{K}([a, x], [a', x']) = k(a, a')k(x, x')$$

Ridge regression solution:

$$\hat{\gamma}(x, a) = \sum_{i=1}^{n} y_i \beta_i(a, x), \quad \beta(a, x) = [K_{AA} \odot K_{XX} + \lambda I]^{-1} K_{Aa} \odot K_{Xx}$$

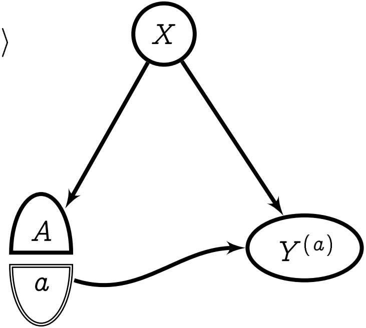
ATE (dose-response curve)

Well-specified setting:

$$\mathbb{E}[Y|a,x] =: \gamma_0(a,x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

ATE as feature space dot product:

$$\begin{aligned} \mathrm{ATE}(a) &= \mathbb{E}[\gamma_0(a, X)] \\ &= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle\right] \end{aligned}$$



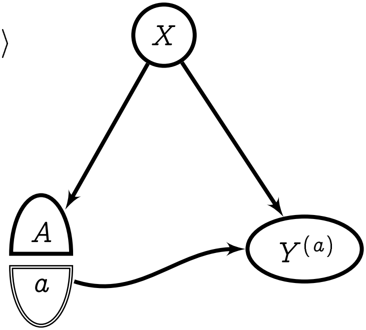
ATE (dose-response curve)

Well-specified setting:

$$\mathbb{E}[Y|a,x] =: \gamma_0(a,x) = \langle \gamma_0, \varphi(a) \otimes \varphi(x) \rangle$$

ATE as feature space dot product:

$$\begin{aligned} \text{ATE}(a) &= \mathbb{E}[\gamma_0(a,X)] \\ &= \mathbb{E}[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle] \\ &= \langle \gamma_0, \varphi(a) \otimes \underbrace{\mu_X}_{\mathbb{E}[\varphi(X)]} \rangle \end{aligned}$$



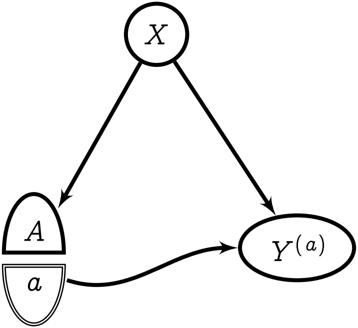
Feature map of probability $P(X)$,

$$\mu_X = [\ldots \mathbb{E}[\varphi_i(X)] \ldots]$$

ATE: example

US job corps: training for disadvantaged youths:

- X: covariate/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employment)



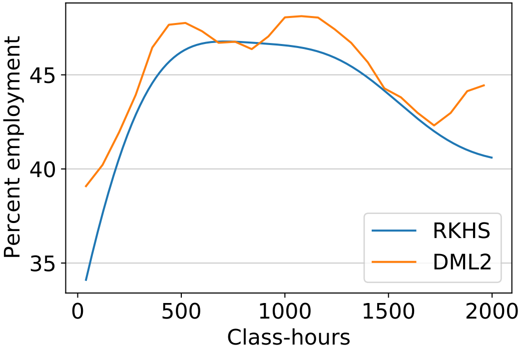
Empirical ATE:

$$\widehat{\text{ATE}}(a) = \hat{\mathbb{E}}\left[\langle \hat{\gamma}_0, \varphi(X) \otimes \varphi(a)\rangle\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1}(K_{Aa} \odot K_{Xx_i})$$

Schochet, Burghardt, and McConnell (2008). Does Job Corps work? Impact findings from the national Job Corps study.

Singh, Xu, G (2022a).

ATE: results



- First 12.5 weeks of classes confer employment gain: from 35% to 47%.
- [RKHS] is our $\widehat{\mathrm{ATE}}(a)$.
- [DML2] Colangelo, Lee (2020), Double debiased machine learning nonparametric inference with continuous treatments.

Singh, Xu, G (2022a)

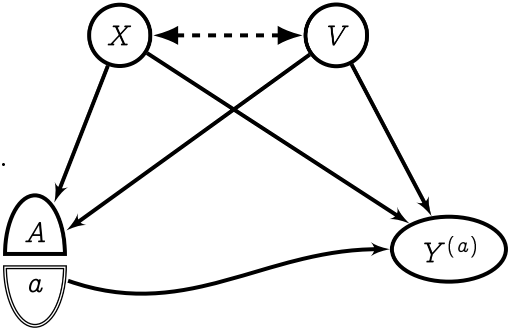
Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[\,Y\,|\,a, x, v\,] =: \gamma_0(a, x, v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle\,.$$

Conditional ATE

$$\text{CATE}(a, v)$$
$$= \mathbb{E}\left[Y^{(a)} \middle|\, V = v \right]$$



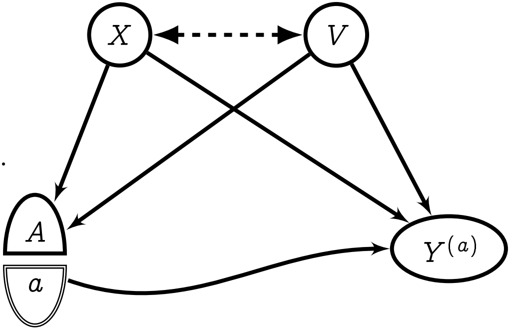
Conditional average treatment effect

Well-specified setting:

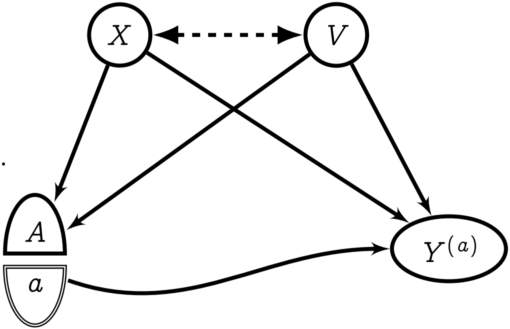
$$\mathbb{E}[Y|a,x,v] =: \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle \,.$$

Conditional ATE

$$\text{CATE}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v\right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v\right]$$



Conditional average treatment effect



Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle \,.$$

Conditional ATE

$$\mathrm{CATE}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} | V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$
$$= ...?$$

How to take conditional expectation?

Density estimation for $p(X|V=v)$? Sample from $p(X|V=v)$?

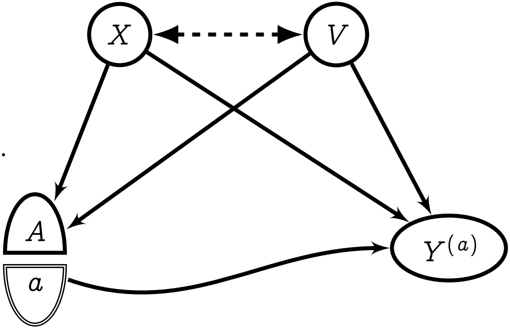
Conditional average treatment effect

Well-specified setting:

$$\mathbb{E}[Y|a,x,v] =: \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle .$$

Conditional ATE

$$\mathrm{CATE}(a,v)$$
$$= \mathbb{E}\left[Y^{(a)} \middle| V = v \right]$$
$$= \mathbb{E}\left[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v \right]$$
$$= \langle \gamma_0, \varphi(a) \otimes \underbrace{\mathbb{E}[\varphi(X)| V = v]}_{\mu_{X|V=v}} \otimes \varphi(v) \rangle$$



Learn conditional mean embedding: $\mu_{X|V=v} := \mathbb{E}_X\left[\varphi(X)| V = v\right]$

Regressing *from* feature space *to* feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X \mid V=v}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing *from* feature space *to* feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\operatorname{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}, \mathcal{H}_{\mathcal{X}})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_{\mathcal{V}} \quad \forall h \in \mathcal{H}_{\mathcal{X}}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing *from* feature space *to* feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\operatorname{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}, \mathcal{H}_{\mathcal{X}})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_{\mathcal{V}} \quad \forall h \in \mathcal{H}_{\mathcal{X}}$$

A Smooth Operator

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing *from* feature space *to* feature space

Our goal: an operator $F_0 : \mathcal{H}_{\mathcal{V}} \rightarrow \mathcal{H}_{\mathcal{X}}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \text{HS}(\mathcal{H}_{\mathcal{V}}, \mathcal{H}_{\mathcal{X}})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|V = v] \in \mathcal{H}_{\mathcal{V}} \quad \forall h \in \mathcal{H}_{\mathcal{X}}$$

Kernel ridge regression from $\varphi(v)$ to *infinite* features $\varphi(x)$:

$$\widehat{F} = \underset{F \in HS}{\text{argmin}} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_{\mathcal{X}}} + \lambda_2 \|F\|^2_{HS}$$

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012). Conditional mean embeddings as regressors.

Grunewalder, G, Shawe-Taylor (2013) Smooth operators.

Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning

Regressing *from* feature space *to* feature space

Our goal: an operator $F_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$F_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$F_0 \in \overline{\text{span}\,\{\varphi(x) \otimes \varphi(v)\}} \iff F_0 \in \text{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

$$\mathbb{E}[h(X)|\, V = v] \in \mathcal{H}_\mathcal{V} \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Kernel ridge regression from $\varphi(v)$ to *infinite* features $\varphi(x)$:

$$\widehat{F} = \underset{F \in HS}{\text{argmin}} \sum_{\ell=1}^{n} \|\varphi(x_\ell) - F\varphi(v_\ell)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda_2 \|F\|^2_{HS}$$

Ridge regression solution:

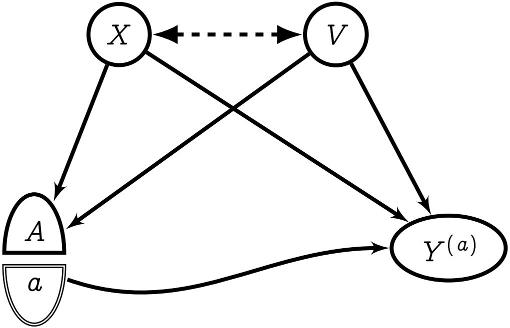
$$\mu_{X|V=v} := \mathbb{E}[\varphi(X)|\, V = v] \approx \widehat{F}\varphi(v) = \sum_{\ell=1}^{n} \varphi(x_\ell)\beta_\ell(v)$$

$$\beta(v) = [K_{VV} + \lambda_2 I]^{-1} k_{Vv}$$

Conditional ATE: example

US job corps:

- X: confounder/context (education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employed)
- V: age



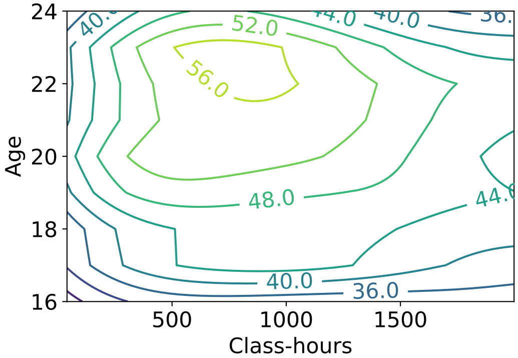
Empirical CATE:

$$\widehat{\text{CATE}}(a, v) = \langle \hat{\gamma}_0, \varphi(a) \otimes \underbrace{\widehat{F}\varphi(v)}_{\widehat{\mathbb{E}}[\varphi(X) \mid V=v]} \otimes \varphi(v) \rangle$$

(with consistency guarantees: see paper!)

Singh, Xu, G (2022a)

Conditional ATE: results



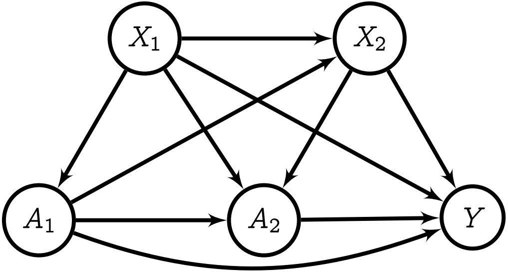
Average percentage employment $Y^{(a)}$ for class hours a, conditioned on age v. Given around 12-14 weeks of classes:

- 16 y/o: employment increases from 28% to at most 36%.
- 22 y/o: percent employment increases from 40% to 56%.

Singh, Xu, G (2022a)

...dynamic treatment effect...

Dynamic treatment effect: sequence A_1, A_2 of treatments.



- potential outcomes $Y^{(a_1)}$, $Y^{(a_2)}$, $Y^{(a_1,a_2)}$,
- counterfactuals $\mathbb{E}\left[Y^{(a_1',a_2')}|A_1 = a_1, A_2 = a_2\right]$...

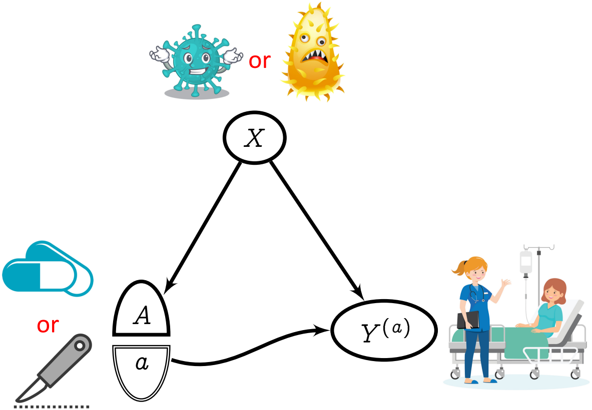
(c.f. the Robins G-formula)

Singh, Xu, G. (2022b) Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects

What if there are hidden confounders?

Reminder: observation vs intervention

Average causal effect (intervention): $\mathbb{E}[Y^{(a)}] = \sum_{x \in \{0,1\}} \mathbb{E}[Y | a, x] p(x)$

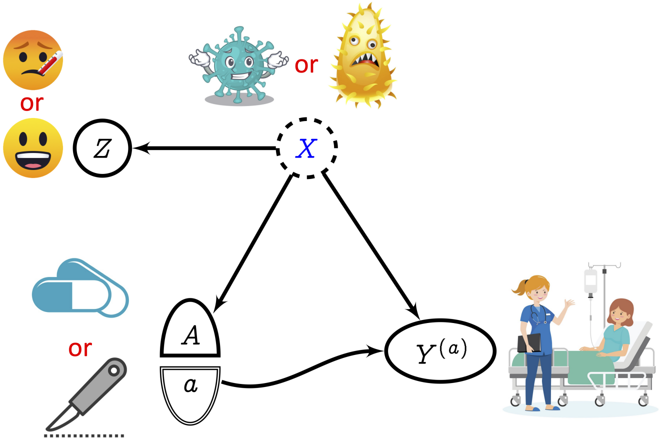


From our *intervention* (making all patients take a treatment):

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

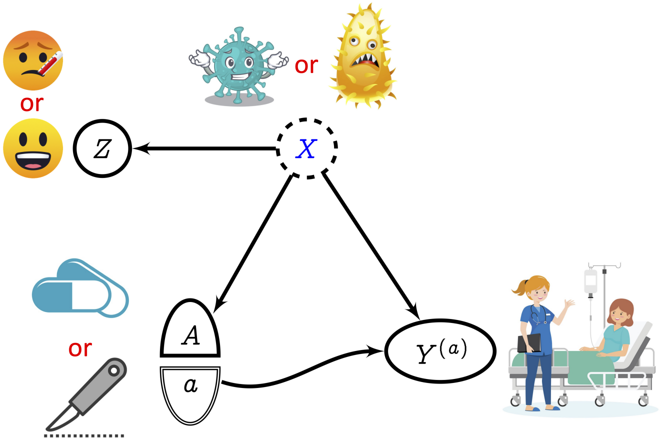
Richardson, Robins (2013), Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

We observe symptom Z, not disease X



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

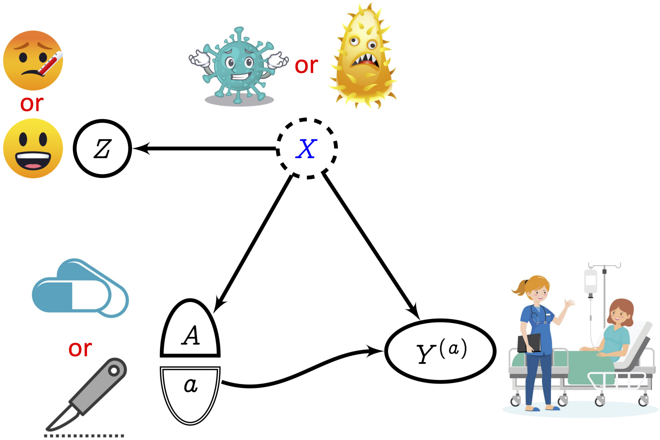
We observe symptom Z, not disease X



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

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

We observe symptom Z, not disease X



Results are very bad:

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

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)

What's new? What is it good for?

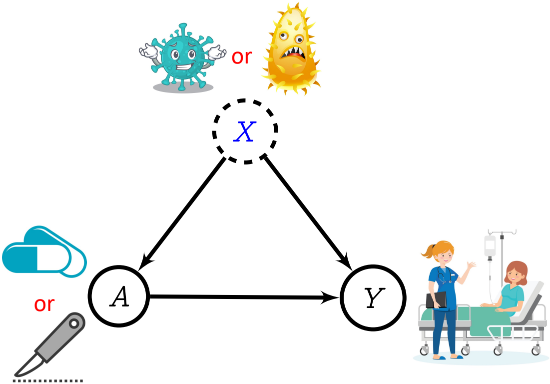
- Treatment A, proxy variables, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations
- Don't ~~meet your heroes~~ model your hidden variables!

Proxy variables: health example

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

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome



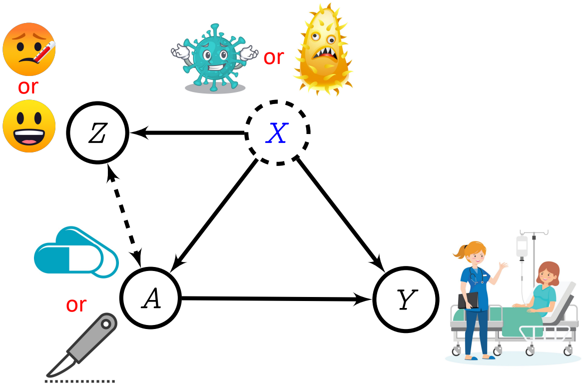
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

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

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms



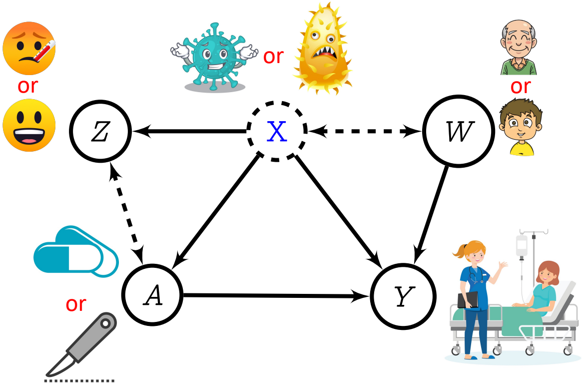
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

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

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms
- W: age



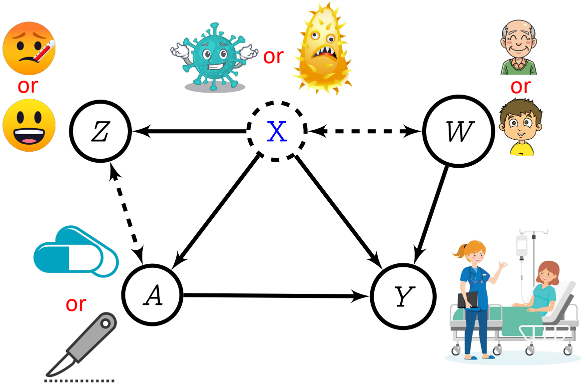
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

Proxy variables: health example

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

The definitions are:

- X: underlying illness severity
- A: treatment
- Y: outcome
- Z: symptoms
- $: W$ age



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

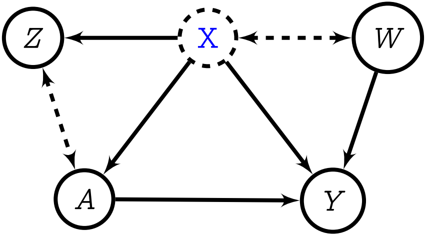
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder.

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

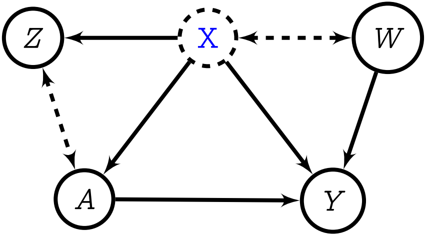


Proxy variables: general setting

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:

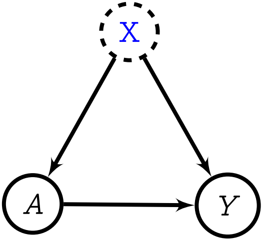
$$W \perp\!\!\!\perp (Z, A) | X$$

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

Why proxy variables? A simple proof

The definitions are:

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



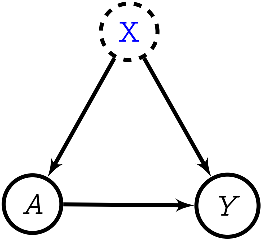
If X were observed,

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

Why proxy variables? A simple proof

The definitions are:

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



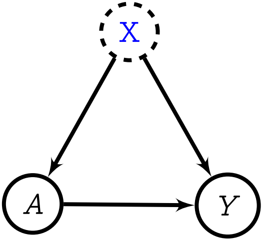
If X were observed,

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

Why proxy variables? A simple proof

The definitions are:

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



If X were observed,

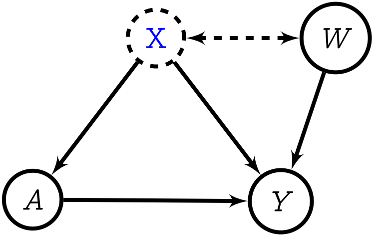
$$\underbrace{P(Y^{(a)})}_{d_y \times 1} := \sum_{i=1}^{d_x} P(Y|x_i, a)P(x_i) = \underbrace{P(Y|X, a)}_{d_y \times d_x}\underbrace{P(X)}_{d_x \times 1}$$

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



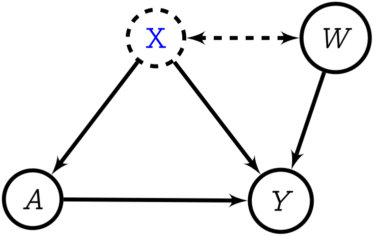
For each a, if we could solve:

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

...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:

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

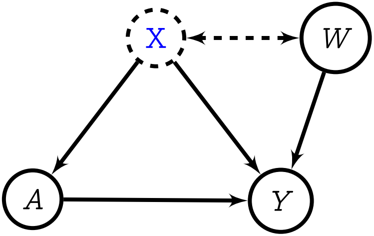
.....then

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

...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:

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

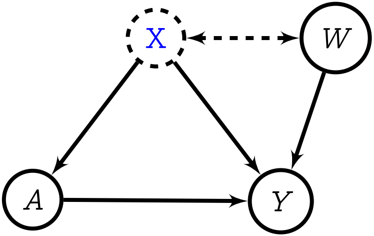
.....then

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

...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:

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

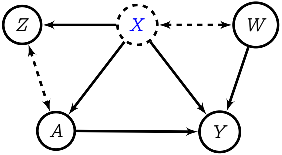
.....then

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

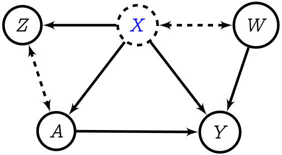
...now project onto $p(X|Z,a)$

From last slide,

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



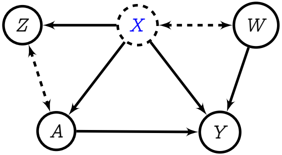
...now project onto $p(X|Z, a)$



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 $p(X|Z,a)$



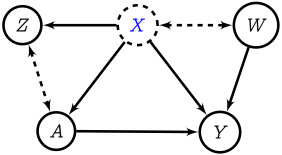
From last slide,

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

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,

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

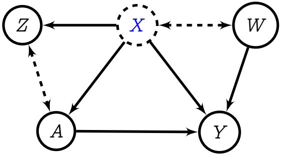
Because $W \perp\!\!\!\perp (Z,A)|X$,

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

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!

Proxy/Negative Control Methods in the Real World

Unobserved confounders: proxy methods

Kernel features (ICML 2021):

arXiv.org > cs > arXiv:2105.04544

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

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/

Unobserved confounders: proxy methods

Kernel features (ICML 2021):

arXiv.org > cs > arXiv:2105.04544

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

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/

One model: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x) = \langle \gamma, \varphi_\theta(x) \rangle_{\mathcal{H}}$$

One model: linear functions of features

All learned functions will take the form:

$$\gamma(x) = \gamma^\top \varphi_\theta(x) = \langle \gamma, \varphi_\theta(x) \rangle_{\mathcal{H}}$$

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

Xu, G., A Neural mean embedding approach for back-door and front-door adjustment. (ICLR 23)

Xu, Chen, Srinivasan, de Freitas, Doucet, G. Learning Deep Features in Instrumental Variable Regression. (ICLR 21)

Xu, Kanagawa, G. "Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation". (NeurIPS 21)

Kernel approach: Infinite dictionaries of fixed kernel features:

$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Kernel is feature dot product.

Singh, Xu, G. Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. (Biometrika, 2023)

Singh, Sahani, G. Kernel Instrumental Variable Regression. (NeurIPS 19)

Mastouri*, Zhu*, Gultchin, Korba, Silva, Kusner, G,† Muandet† (2021); Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction (ICML21)

Model fitting: *neural* ridge regression

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

$$\hat{\gamma} = \arg\min_{\gamma \in \mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi_\theta(x_i) \rangle_{\mathcal{H}}\right)^2 + \lambda \|\gamma\|_{\mathcal{H}}^2 \right) \quad (1)$$

Model fitting: *neural* ridge regression

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

$$\hat{\gamma} \quad = \quad \arg\min_{\gamma \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle \gamma, \varphi_\theta(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|_{\mathcal{H}}^2 \right) \qquad (1)$$

Solution for linear final layer γ:

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

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

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

Model fitting: *neural* ridge regression

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

$$\hat{\gamma} = \arg\min_{\gamma\in\mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle \gamma, \varphi_\theta(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|\gamma\|^2_{\mathcal{H}} \right) \qquad (1)$$

Solution for linear final layer γ:

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

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

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

How to solve for θ:

Substitute $\hat{\gamma}$ into (1), backprop through Cholesky for θ.

Model fitting: *neural* ridge regression

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

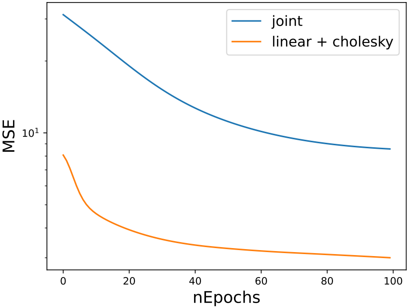
$$\hat{\gamma} \;=\; \arg\min_{\gamma\in\mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle \gamma, \varphi_\theta(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|\gamma\|^2_{\mathcal{H}} \right) \qquad (1)$$

Solution for linear final layer γ:

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

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

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



MNIST, 4 layer FF, sigmoid, fully connected

How to solve for θ:

Substitute $\hat{\gamma}$ into (1), backprop through Cholesky for θ.

Proxy methods, general domains

If X were observed, we would write (average treatment effect)

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

....but we do not observe X.

Proxy methods, general domains

If X were observed, we would write (average treatment effect)

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

....but we do not observe X.

Main theorem: Assume we solved for link function:

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

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

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Proxy methods, general domains

If X were observed, we would write (average treatment effect)

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

....but we do not observe X.

Main theorem: Assume we solved for link function:

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

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

Average treatment effect via $p(w)$:

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

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Proxy methods, general domains

If X were observed, we would write (average treatment effect)

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

....but we do not observe X.

Main theorem: Assume we solved for link function:

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

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

Average treatment effect via $p(w)$:

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

Challenge: need to parametrize and solve for h_y

(Fredholm equation of first kind: existence of solution requires identifiability conditions)

Link function NN parametrization

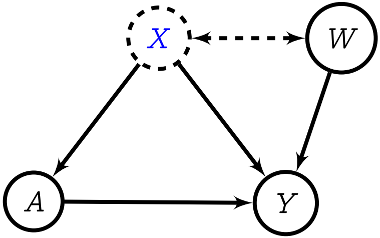
The link function is a function of two arguments

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

Assume we have:

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

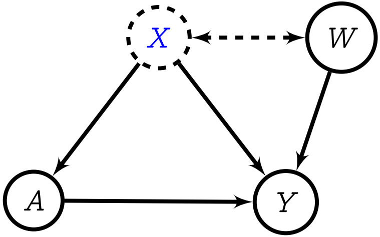
(argument of feature map indicates feature space)



Link function NN parametrization

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 slides)!

Ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Why?

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

Ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Why?

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

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

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

Ridge regression for $h_y(w, a)$

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Why?

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

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

...and by the proxy model above,

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

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

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

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

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

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

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a)p(W|a, Z)dw$$

Ridge regression solution: proxy loss

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

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

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).

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

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

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:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Recall link function

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

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

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

Goal:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Recall link function

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

(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:

$$\mathbb{E}(Y|a, Z) = \int_w h_y(W, a) p(W|a, Z) dw$$

Ridge regression solution: proxy loss

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

Recall link function

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

Ridge regression (again!)

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

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

Challenge: how to learn θ?

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

Challenge: how to learn θ?

From Stage 2 regression?

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

Challenge: how to learn θ?

From Stage 2 regression?

...which requires $\mathbb{E}_{W|a,z}\varphi_\theta(W)$ from Stage 1 regression

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

Challenge: how to learn θ?

From Stage 2 regression?

...which requires $\mathbb{E}_{W|a,z}\varphi_\theta(W)$ from Stage 1 regression

...which requires $\varphi_\theta(W)$... which requires θ...

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

Primary regression: learn NN features $\varphi_\theta(W), \varphi_\xi(A)$ and linear layer γ to obtain Y with RR loss:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

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

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

with RR loss

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

Challenge: how to learn θ?

From Stage 2 regression?

...which requires $\mathbb{E}_{W|a,z}\varphi_\theta(W)$ from Stage 1 regression

...which requires $\varphi_\theta(W)$... which requires θ...

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

Learning the auxiliary task

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

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

with RR loss

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

Learning the auxiliary task

Auxiliary regression: learn NN features $\phi_\zeta(Z)$ 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 ϕ_θ, ϕ_ζ:

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

$$C^{(\zeta)}_{AZ} = \mathbb{E}[\phi_\zeta(A,Z)\phi_\zeta^\top(A,Z)]$$

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

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2 \|\gamma\|^2$$

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right]\otimes\varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right]\otimes\varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ,θ,ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$
- Substitute $\hat{\gamma}$ into Stage 2, gradient steps on θ,ξ
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$
- Substitute $\hat{\gamma}$ into Stage 2, gradient steps on θ, ξ
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.
  - Iterate between θ, ξ and ζ

Xu, Kanagawa, G. (2021).

Final algorithm

Primary regression:

$$\mathbb{E}_{Y,A,Z}\left(Y - \gamma^\top\left(\mathbb{E}_{W|A,Z}\left[\varphi_\theta(W)\right] \otimes \varphi_\xi(A)\right)\right)^2 + \lambda_2\|\gamma\|^2$$

Auxiliary regression: NN params ζ and $\hat{F}_{\theta,\zeta}$:

$$\mathbb{E}_{W|A,Z}[\varphi_\theta(W)] \approx \hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$$

Solution procedure: for γ, θ, ξ:

- Get $\hat{\gamma}$ in closed form as function of $\hat{F}_{\theta,\zeta}\phi_\zeta(A,Z)$ and $\varphi_\xi(A)$
- Substitute $\hat{\gamma}$ into Stage 2, gradient steps on θ, ξ
  - $\hat{F}_{\theta,\zeta}$ remains optimal wrt current φ_θ.
  - Iterate between θ, ξ and ζ

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

$$\mathbb{E}(Y|a,Z) = \int_w h_y(W,a)p(W|a,Z)dw$$

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

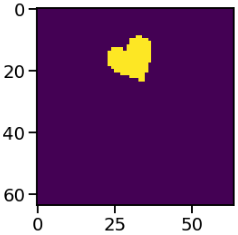
Xu, Kanagawa, G. (2021).

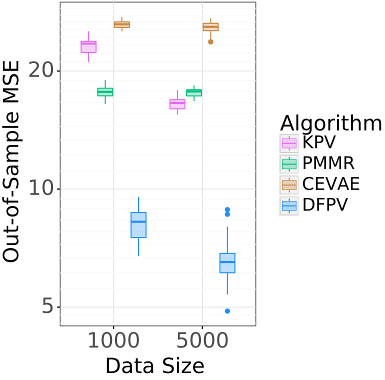
Experiments

Synthetic experiment, adaptive neural net features

dSprite example:

- $X = \{\texttt{scale}, \texttt{rotation}, \texttt{posX}, \texttt{posY}\}$
- Treatment A is the image generated (with Gaussian noise)
- Outcome Y is quadratic function of A with multiplicative confounding by posY.
- $Z = \{\texttt{scale}, \texttt{rotation}, \texttt{posX}\}$, W = noisy image sharing posY
- Comparison with CEVAE (Louzios et al. 2017)



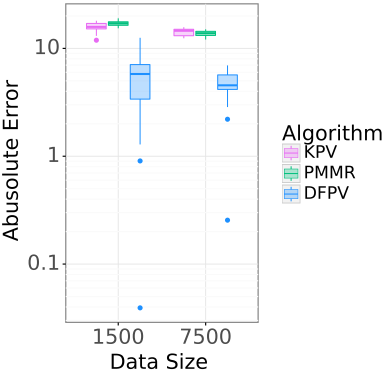


Louizos, Shalit, Mooij, Sontag, Zemel, Welling, Causal Effect Inference with Deep Latent-Variable Models (2017)

Confounded offline policy evaluation

Synthetic dataset, demand prediction for flight purchase.

- Treatment A is ticket price.
- Policy $A \sim \pi(Z)$ depends on fuel price.

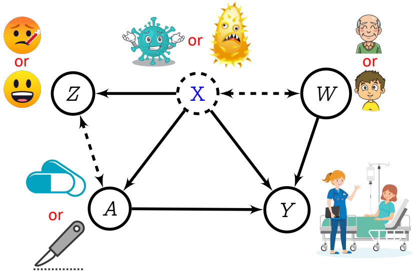


Conclusion

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)

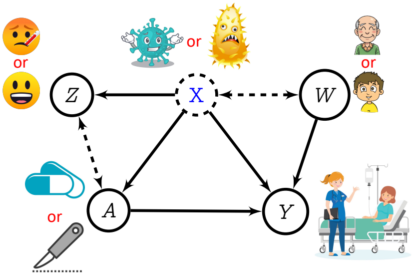


Conclusion

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/

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Questions?

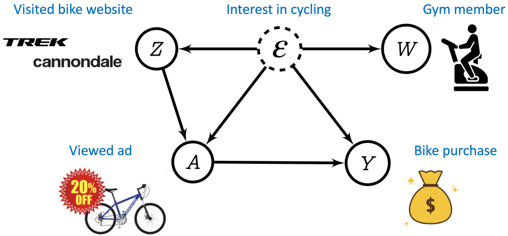

THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Web ads example

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

The definitions are:

- ε: "interest in cycling"
- A: bike ad on browser
- Y: purchase
- Z: visit to bike website $\implies$ cookies
- W membership of gym



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

Tennenholtz, Mannor, Shalit (2020), OPE in Partially Observed Environments.

Uehara, Sekhari, Lee, Kallus, Sun (2022) Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a, \varepsilon) p(\varepsilon) d\varepsilon.$$

....but we do not observe ε.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)d\varepsilon.$$

....but we do not observe ε.

Main theorem: Assume we solved:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

Both $p(y|a,z)$ and $p(w|a,z)$ are in terms of observed quantities.

Main theorem

If ε were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)d\varepsilon.$$

....but we do not observe ε.

Main theorem: Assume we solved:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

Both $p(y|a,z)$ and $p(w|a,z)$ are in terms of observed quantities.

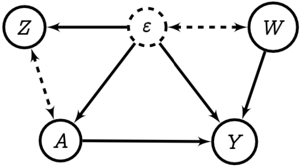
Average treatment effect via $p(w)$:

$$p(y^{(a)}) = \int h_y(a,w)p(w)dw$$

Proof (1)

Because $W \perp\!\!\!\perp (Z, A) | \varepsilon$, we have

$$p(w|a, z) = \int p(w|\varepsilon)p(\varepsilon|a, z)d\varepsilon$$



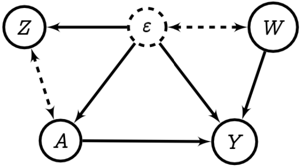
Proof (1)

Because $W \perp\!\!\!\perp (Z, A) | \varepsilon$, we have

$$p(w|a, z) = \int p(w|\varepsilon)p(\varepsilon|a, z)d\varepsilon$$

Because $Y \perp\!\!\!\perp Z | (A, \varepsilon)$ we have

$$p(y|a, z) = \int p(y|a, \varepsilon)p(\varepsilon|a, z)d\varepsilon$$



Proof (3)

Given the solution h_y to:

$$p(y|a, z) = \int h_y(w, a) p(w|a, z) dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

Proof (3)

Given the solution h_y to:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

$$\int p(y|a,\varepsilon)p(\varepsilon|a,z)d\varepsilon = \int h_y(w,a) \int p(w|\varepsilon)p(\varepsilon|a,z)d\varepsilon\, dw$$

Proof (3)

Given the solution h_y to:

$$p(y|a,z) = \int h_y(w,a)p(w|a,z)dw$$

(well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

$$\int p(y|a,\varepsilon)p(\varepsilon|a,z)d\varepsilon = \int h_y(w,a)\int p(w|\varepsilon)p(\varepsilon|a,z)d\varepsilon\, dw$$

This implies:

$$p(y|a,\varepsilon) = \int h_y(w,a)p(w|\varepsilon)dw$$

under identifiability condition

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

Proof (4)

From last slide,

$$p(y|a,\varepsilon) = \int h_y(w,a)p(w|\varepsilon)dw$$

Thus

$$p(y|do(a)) = \int_u p(y|a,\varepsilon)p(\varepsilon)du$$

Proof (4)

From last slide,

$$p(y|a, \varepsilon) = \int h_y(w, a) p(w|\varepsilon) dw$$

Thus

$$\begin{aligned} p(y|do(a)) &= \int_u p(y|a, \varepsilon) p(\varepsilon) du \\ &= \int_u \left[\int h_y(w, a) p(w|\varepsilon) dw \right] p(\varepsilon) d\varepsilon \end{aligned}$$

Proof (4)

From last slide,

$$p(y|a, \varepsilon) = \int h_y(w, a) p(w|\varepsilon) dw$$

Thus

$$\begin{aligned} p(y|do(a)) &= \int_u p(y|a, \varepsilon) p(\varepsilon) du \\ &= \int_u \left[\int h_y(w, a) p(w|\varepsilon) dw \right] p(\varepsilon) d\varepsilon \\ &= \int h_y(w, a) p(w) dw \end{aligned}$$

How not to do 2SLS for proxy methods

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

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

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in HS} \mathbb{E}_{w,a,z} \|\phi(w) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

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

Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W|a,z} \otimes \phi(a) \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in HS} \mathbb{E}_{w,a,z} \|\phi(w) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

Average treatment effect estimate:

$$\mathbb{E}_y(y|do(a)) = \langle h_{\lambda_2}, \phi(a) \otimes \mu_W \rangle,$$

where $\mu_W = \mathbb{E}_W \phi(W)$

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

How *not* to do it

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W,A|a,z} \right\rangle \right)^2 + \lambda_2 \|h\|^2_{\mathcal{H}}$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in \mathcal{G}} \mathbb{E}_{w,a,z} \|\phi(w) \otimes \phi(a) - F[\phi(a) \otimes \phi(z)]\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1 \|F\|^2_{HS}$$

which gives us

$$\mu_{W,A|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

How *not* to do it

Stage 2: minimize

$$h_{\lambda_2} = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left(y - \left\langle h, \mu_{W,A|a,z} \right\rangle \right)^2 + \lambda_2 \|h\|_{\mathcal{H}}^2$$

which is conditional feature mean implementation of

$$p(y|a,z) = \int h_y(w,a) p(w|a,z) dw$$

Stage 1: ridge regression

$$F_{\lambda_1} = \arg\min_{F \in \mathcal{G}} \mathbb{E}_{w,a,z} \|\phi(w) \otimes \phi(a) - F[\phi(a) \otimes \phi(z)]\|_{\mathcal{H}_{\mathcal{W}}}^2 + \lambda_1 \|F\|_{HS}^2$$

which gives us

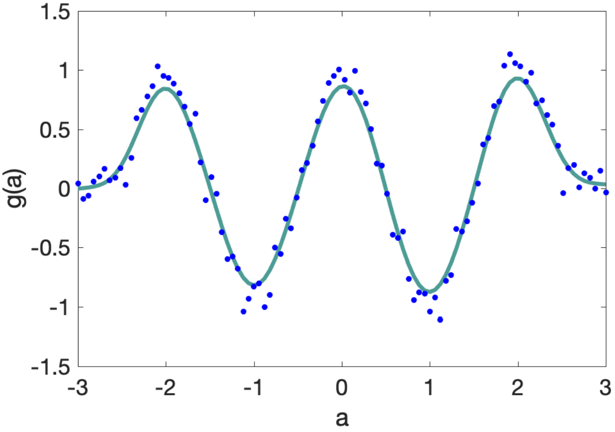
$$\mu_{W,A|a,z} = F_{\lambda_1}[\phi(a) \otimes \phi(z)]$$

Problem: ridge regressing from $\phi(a)$ to $\phi(a)$.

Theoretical issue: $\mathcal{I}_{\mathcal{H}_{\mathcal{A}}}$ is not Hilbert-Schmidt so consistency of F not established.

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

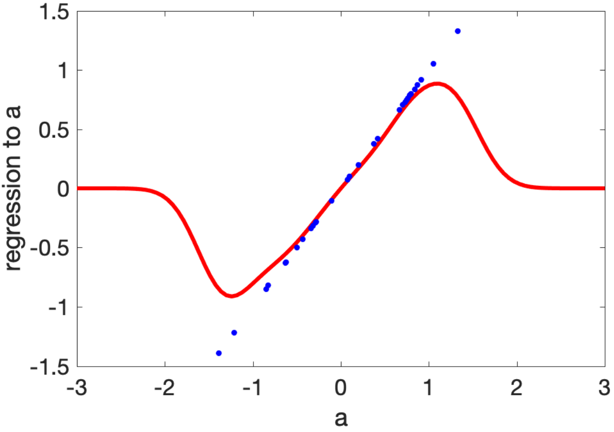
$a \sim \mathcal{N}(0, \sigma^2)$.

Stage 2:

$a \sim \mathcal{U}[-3, 3]$.

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

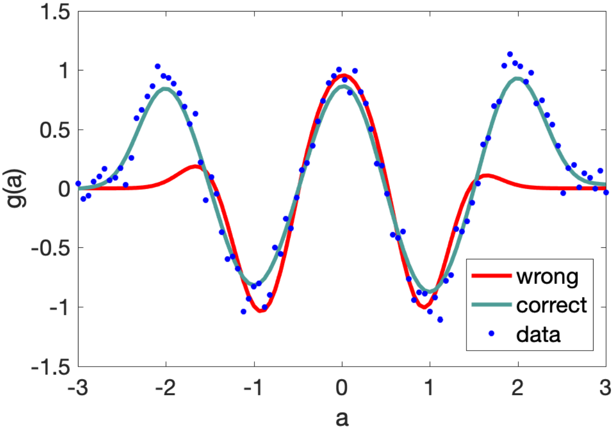
$$a \sim \mathcal{N}(0, \sigma^2).$$

Stage 2:

$$a \sim \mathcal{U}[-3, 3].$$

Demo: bias introduced by stage 1 RR

Implementation issue: this can introduce unnecessary bias.



Stage 1:

$$a \sim \mathcal{N}(0, \sigma^2).$$

Stage 2:

$$a \sim \mathcal{U}[-3, 3].$$

Failures of identifiability assumptions (1)

Recall (one of the) identifiability assumptions:

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

For conciseness, assume conditioning on some a.

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

$$g(\varepsilon) = \tilde{g}(\varepsilon) - \mathbb{E}_{\varepsilon}\tilde{g}(\varepsilon)$$
$$\mathbb{E}(g(\varepsilon)|Z) = \mathbb{E}g(\varepsilon) = 0.$$

Failures of identifiability assumptions (2)

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

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

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

$$\begin{aligned}
&\int_{-\infty}^{\infty} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon \\
&= \int_{-\infty}^{0} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon + \int_{0}^{\infty} g(\varepsilon)p(\varepsilon|z)\,d\varepsilon \\
&= 0.
\end{aligned}$$

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