Causal modelling with neural and kernel feature embeddings: treatment effects, counterfactuals, and proxies

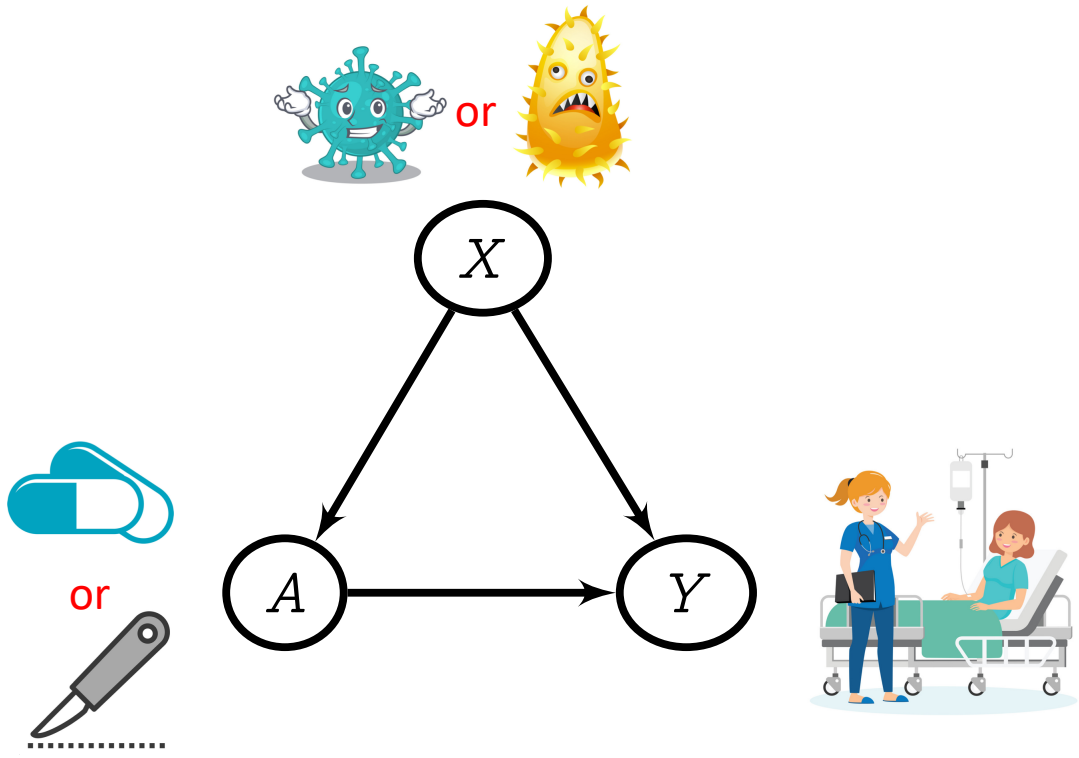
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MIT, 2023

Observation vs intervention

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

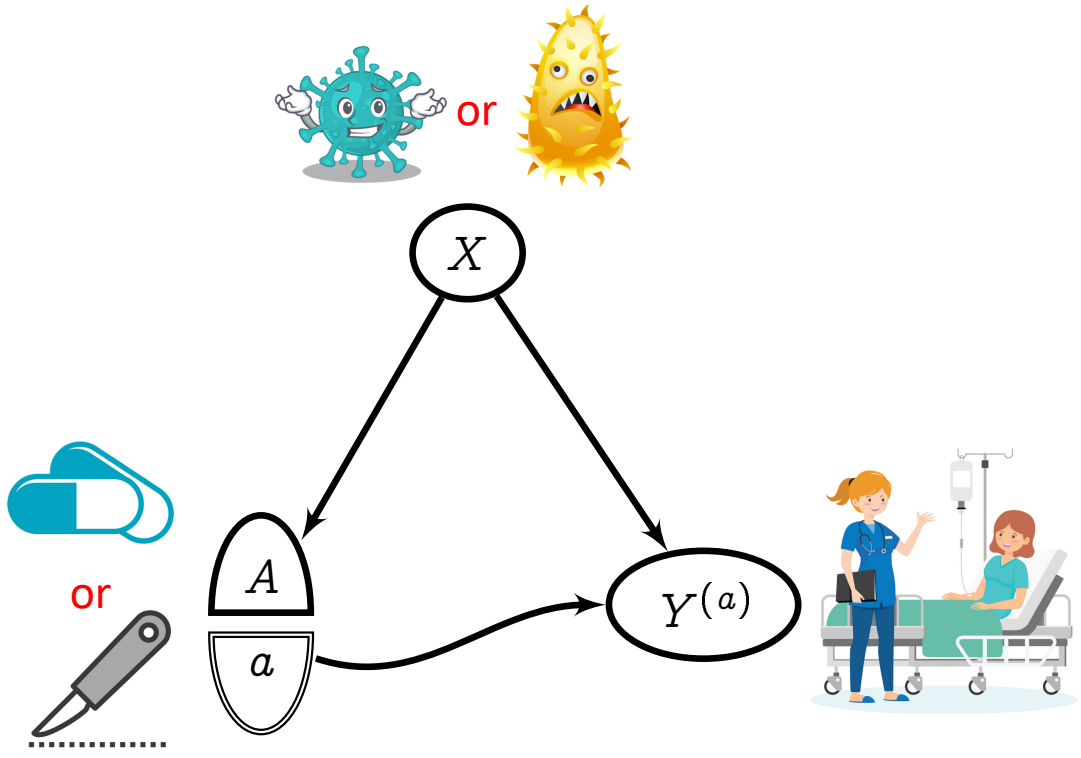


From our observations of historical hospital data:

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

Observation vs intervention

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

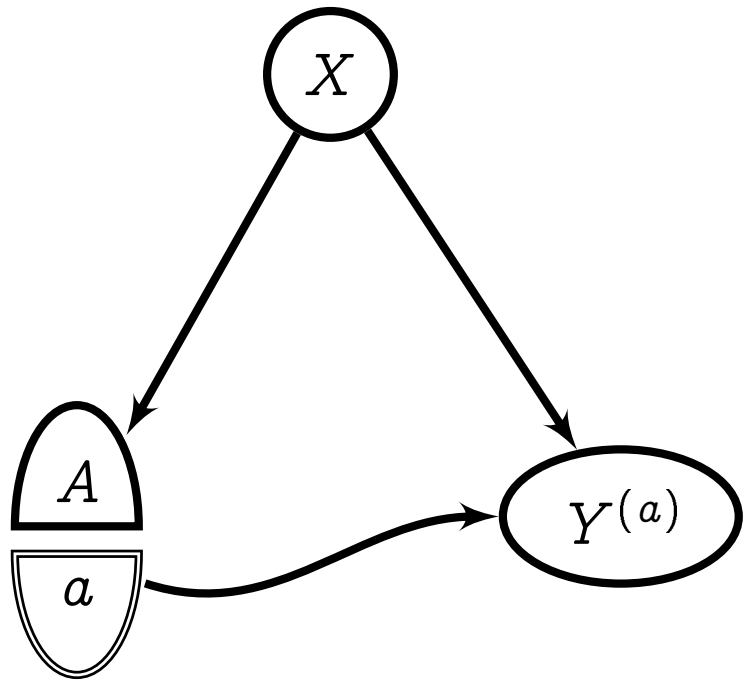


From our intervention (making all patients take a treatment):

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

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

Questions we will solve



Outline

Talk structure:

- Average treatment effect (ATE)
  - ...via kernel mean embedding (marginalization)
- Conditional average treatment effect (CATE)
  - via kernel conditional mean embedding
- Proxy methods
  - ...when covariates are hidden
  - ...causal representation learning via neural conditional mean embedding

Advantages of the approach:

- Treatment A, covariates X, etc can be multivariate, complicated...
- Simple, robust implementation;
- Strong statistical guarantees under general smoothness assumptions (kernel)

Works for kernel or adaptive neural net features!

Key requirement: linear functions of features

All learned functions will take the form:

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

Option 1: Finite dictionaries of learned neural net features (linear final layer)

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

Xu, G., "A Neural mean embedding approach for back-door and front-door adjustment (ICLR23)

Option 2: Infinite dictionaries of fixed kernel features:

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

Kernel is feature dot product.

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

Singh, Xu, G, (2022a) Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves (Biometrika, in revision)

Key building block: 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 as feature dot product:

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

Key building block: ridge regression

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

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

Kernel as feature dot product:

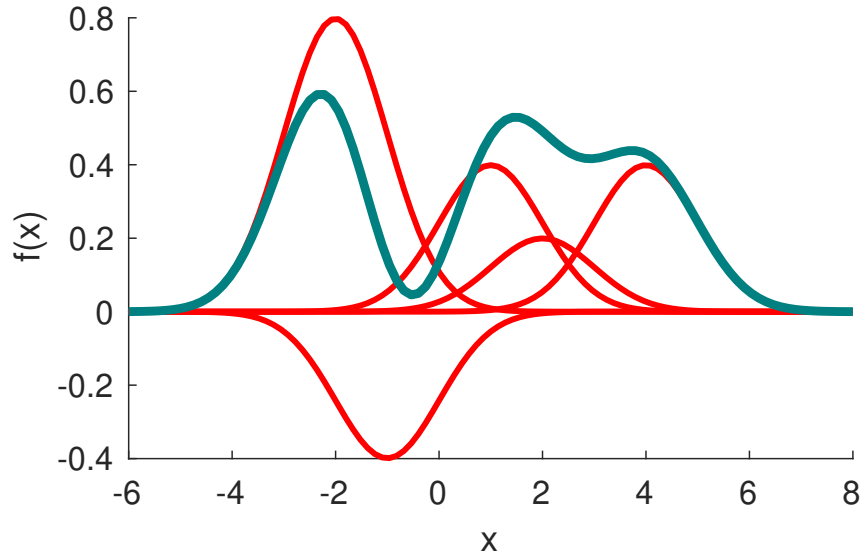
$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Solution at x:

$$\hat{\gamma}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

$$\alpha = (K_{XX} + \lambda I)^{-1} Y$$

$$(K_{XX})_{ij} = k(x_i, x_j),$$



Key building block: 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 as feature dot product:

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

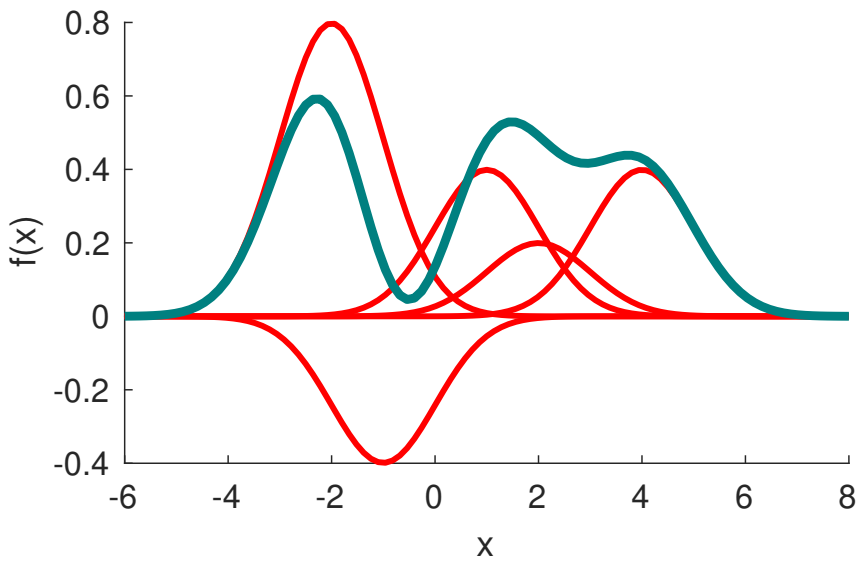
Solution at x (as weighted sum of y)

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

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

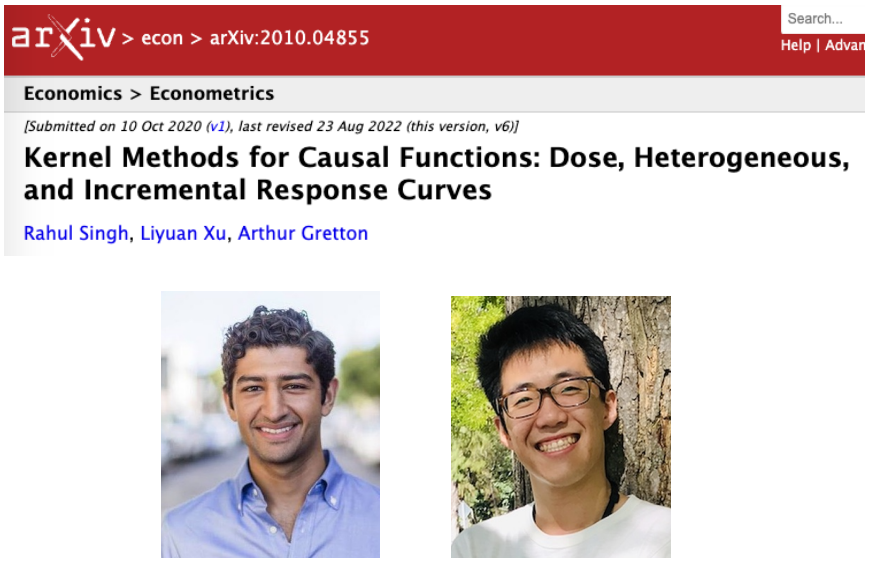
$$(K_{XX})_{ij} = k(x_i, x_j)$$

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



Observed covariates: (conditional) ATE, ATT

Kernel features (in revision, Biometrika):



NN features (ICLR 2023):

Average treatment effect

Potential outcome (intervention):

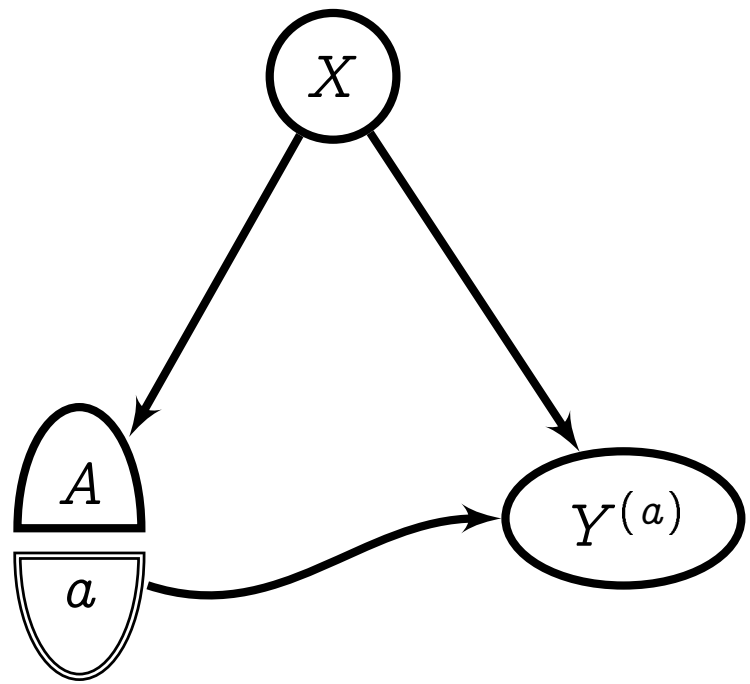
$$\mathbb{E}(Y^{(a)}) = \int \mathbb{E}(y|a,x)\,dp(x)$$

(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

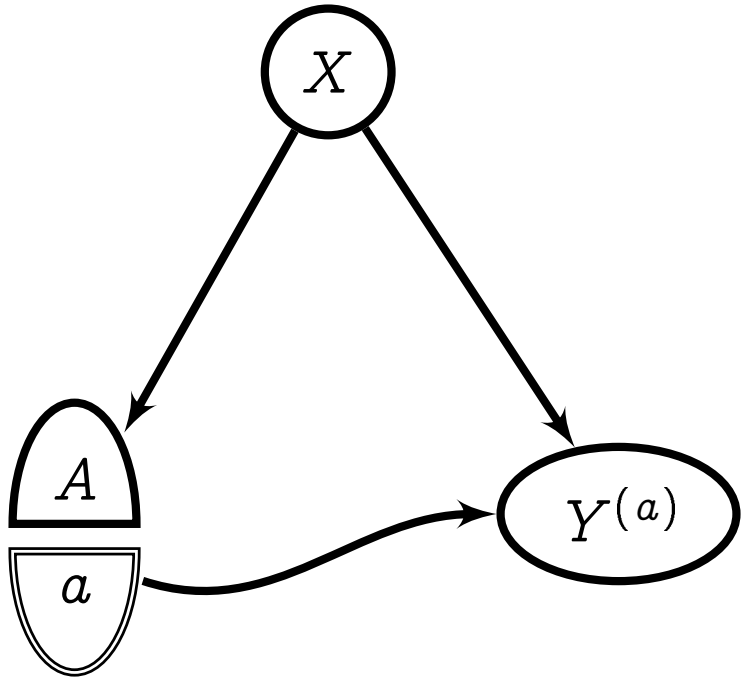
We may predict expected outcome from two inputs

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

Assume we have:

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

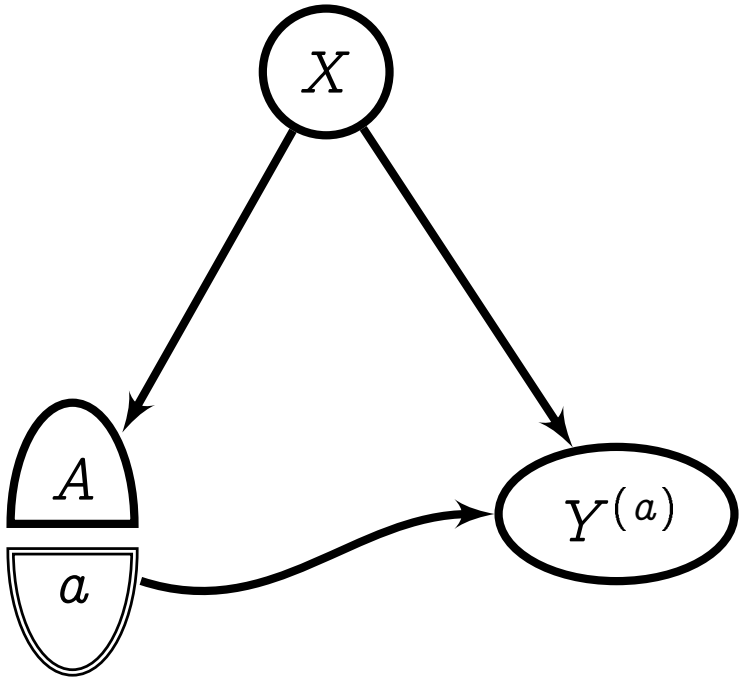
(argument of kernel/feature map indicates feature space)



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)

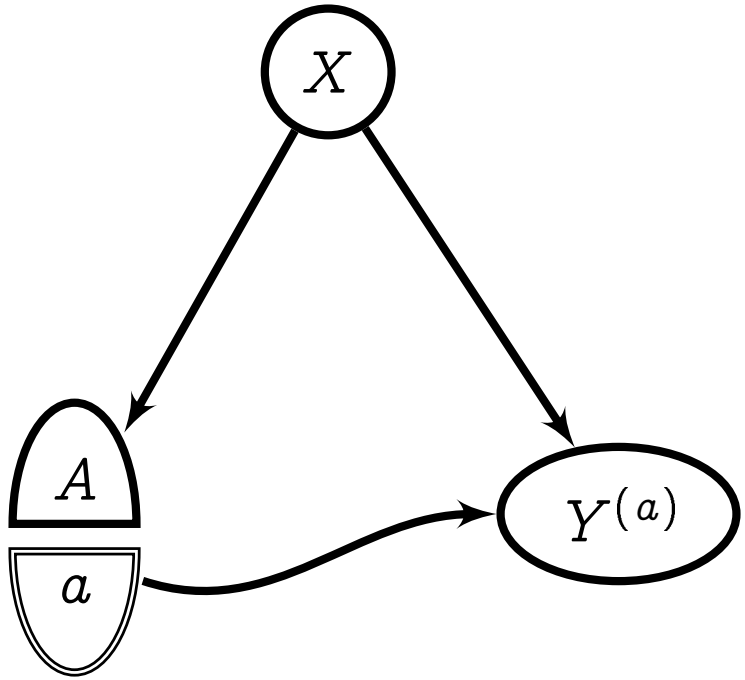
We use outer product of features ($\implies$ product of kernels):

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

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:

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

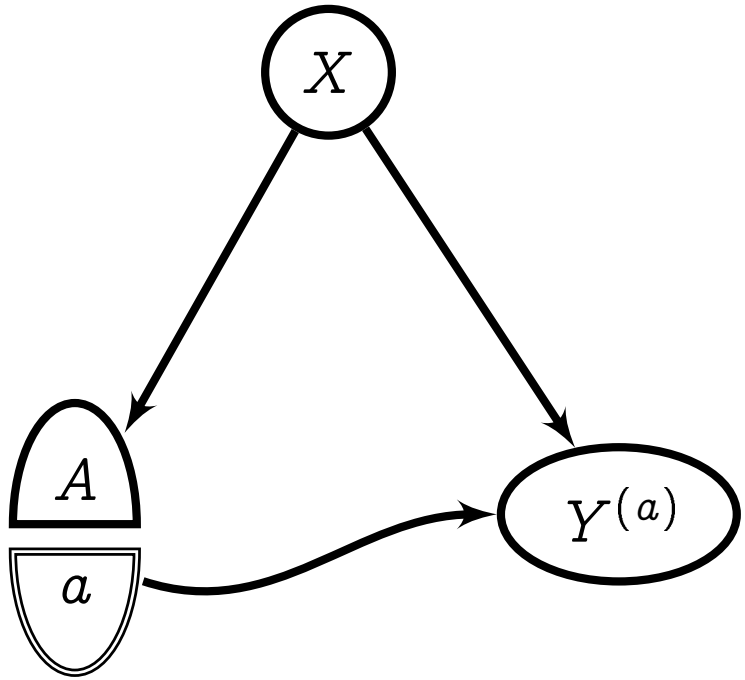
ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

$$\begin{aligned}\theta_0^{\mathrm{ATE}}(a) &= \mathbb{E}_P[\gamma_0(a, X)] \\ &= \mathbb{E}_P \langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle\end{aligned}$$



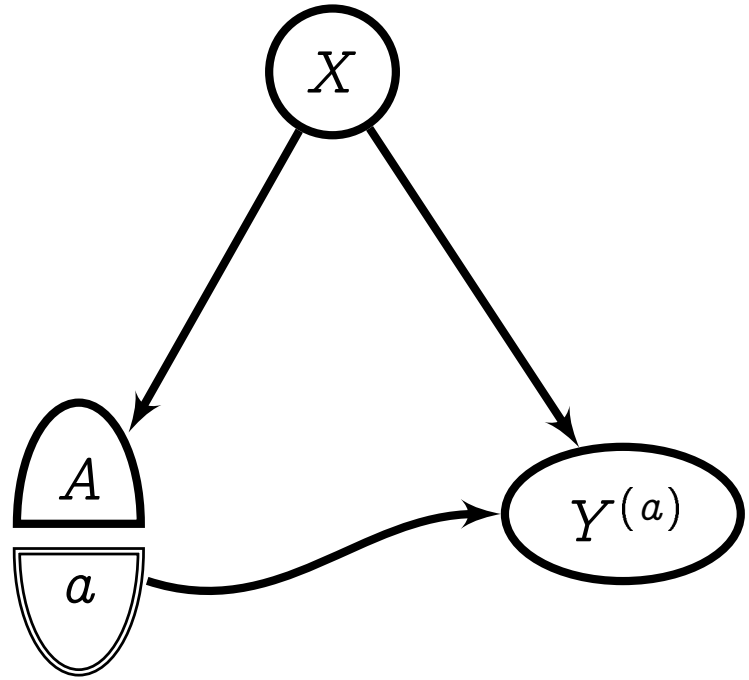
ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

$$\begin{aligned}\theta_0^{\mathrm{ATE}}(a) &= \mathbb{E}_P[\gamma_0(a, X)] \\ &= \mathbb{E}_P \left\langle \gamma_0, \varphi(a) \otimes \varphi(X) \right\rangle \\ &= \left\langle \gamma_0, \underbrace{\mu_P}_{\mathbb{E}_P \varphi(X)} \otimes \varphi(a) \right\rangle\end{aligned}$$



Feature map of probability P,

$$\mu_P = [\ldots \mathbb{E}_P \left[\varphi_i(X)\right] \ldots]$$

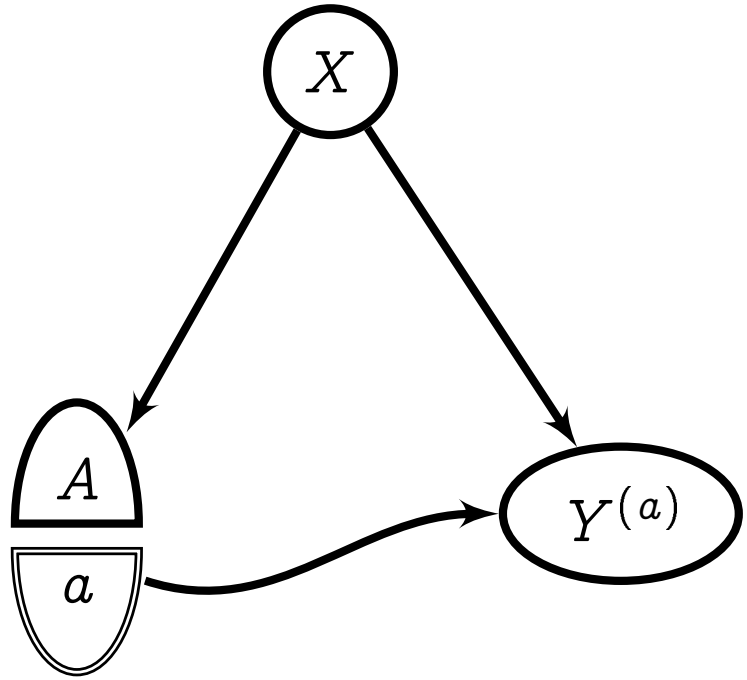
ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

$$\begin{aligned}
\theta_0^{\mathrm{ATE}}(a) &= \mathbb{E}_P[\gamma_0(a, X)] \\
&= \mathbb{E}_P \left\langle \gamma_0, \varphi(a) \otimes \varphi(X) \right\rangle \\
&= \Big\langle \gamma_0, \underbrace{\mu_P}_{\mathbb{E}_P \varphi(X)} \otimes \varphi(a) \Big\rangle
\end{aligned}$$



For characteristic kernels, μ_P is injective.

Consistency: $\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}} = O_P(n^{-1/2})$

ATE: empirical estimate and consistency

Empirical estimate of ATE:

$$\hat{\theta}^{\text{ATE}}(a) = \frac{1}{n}\sum_{i=1}^{n} Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1}(K_{Aa} \odot K_{Xx_i})$$

Singh, Xu, G (2022a), Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves.

ATE: empirical estimate and consistency

Empirical estimate of ATE:

$$\hat{\theta}^{\text{ATE}}(a) = \frac{1}{n}\sum_{i=1}^{n} Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1}(K_{Aa} \odot K_{Xx_i})$$

Consistency:

$$\left\|\hat{\theta}^{\text{ATE}} - \theta_o^{\text{ATE}}\right\|_\infty = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1/b}}\right)$$

Follows from consistency of $\hat{\mu}_P$ and $\hat{\gamma}$, under:

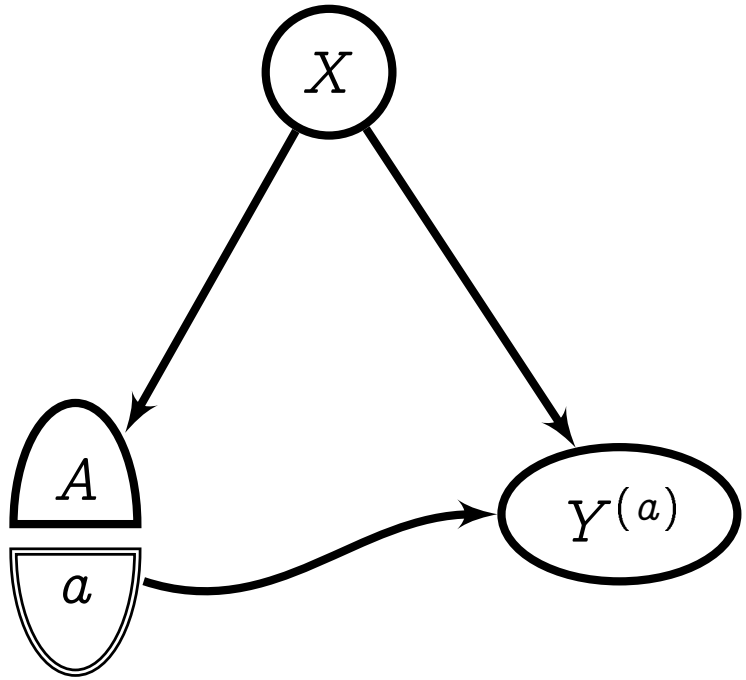
- smoothness assumption $\gamma_0 \in \mathcal{H}^c$, $c \in (1, 2]$
- eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$.

Singh, Xu, G (2022a), Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves.

ATE: example

US job corps: training for disadvantaged youths:

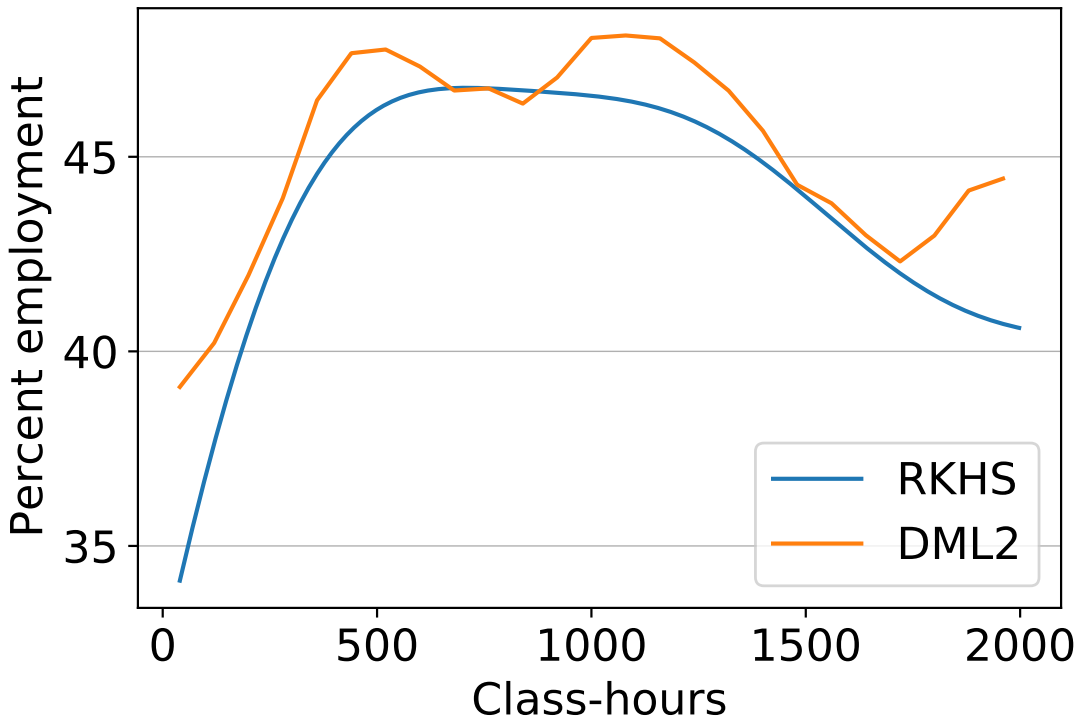
- X: covariate/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employment)



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 $\hat{\theta}^{\mathrm{ATE}}(a)$
- [DML2] Colangelo, Lee (2020), Double debiased machine learning nonparametric inference with continuous treatments.

Singh, Xu, G (2022a)

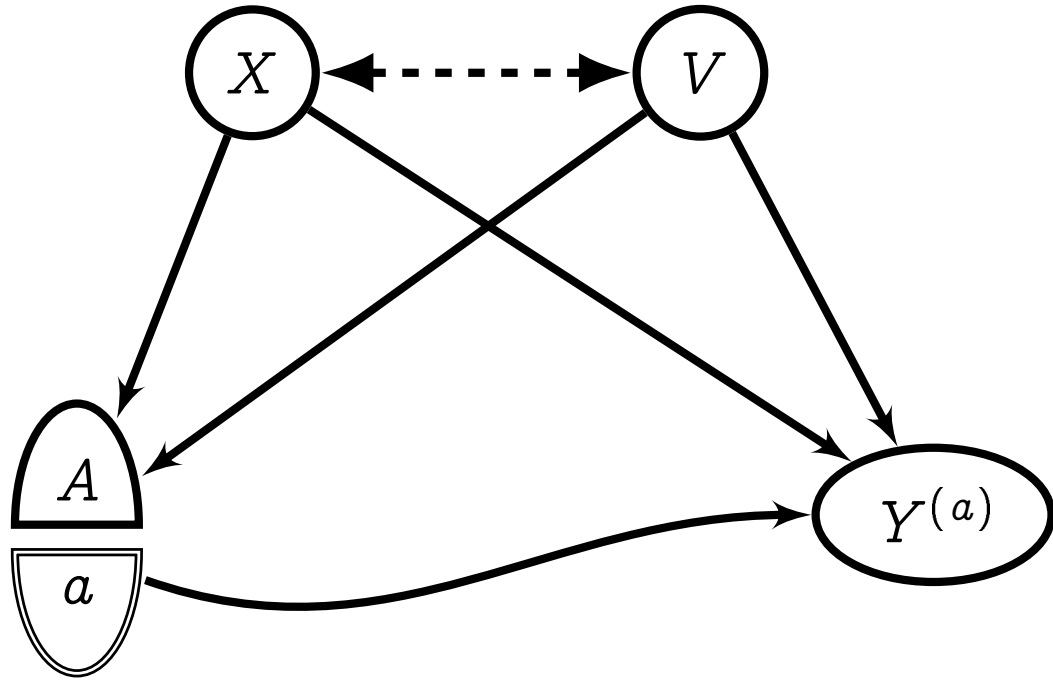
Conditional average treatment effect

Learned conditional mean:

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

Conditional ATE

$$\theta_o^{\mathrm{CATE}}(a,v)$$
$$= \mathbb{E}(\,Y^{(a)}\,|\,V = v)$$



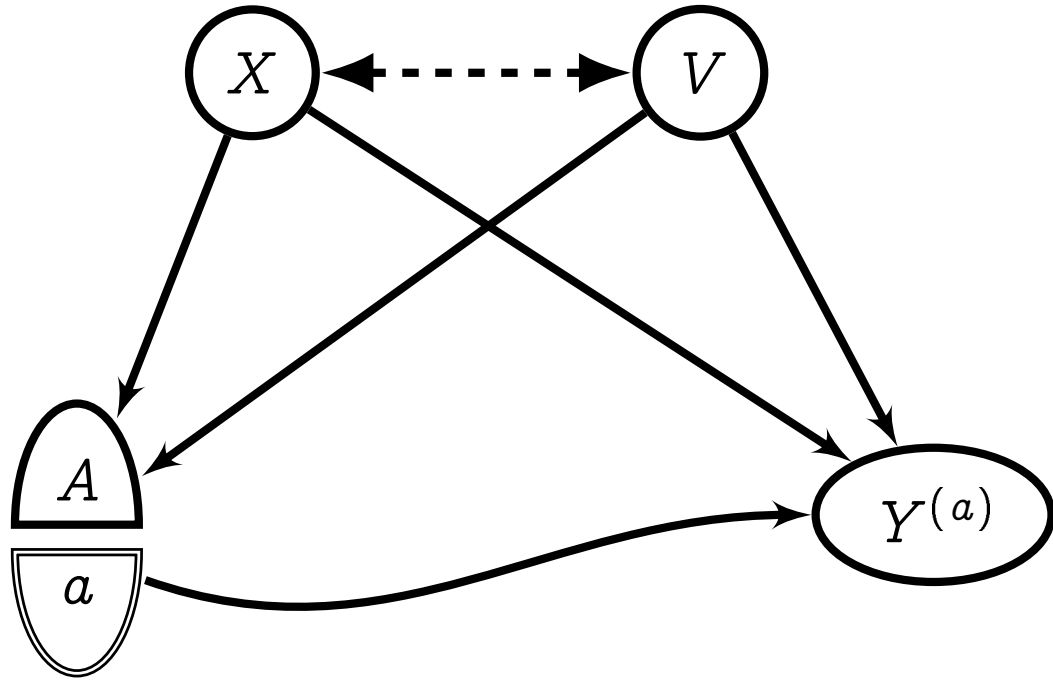
Conditional average treatment effect

Learned conditional mean:

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

Conditional ATE

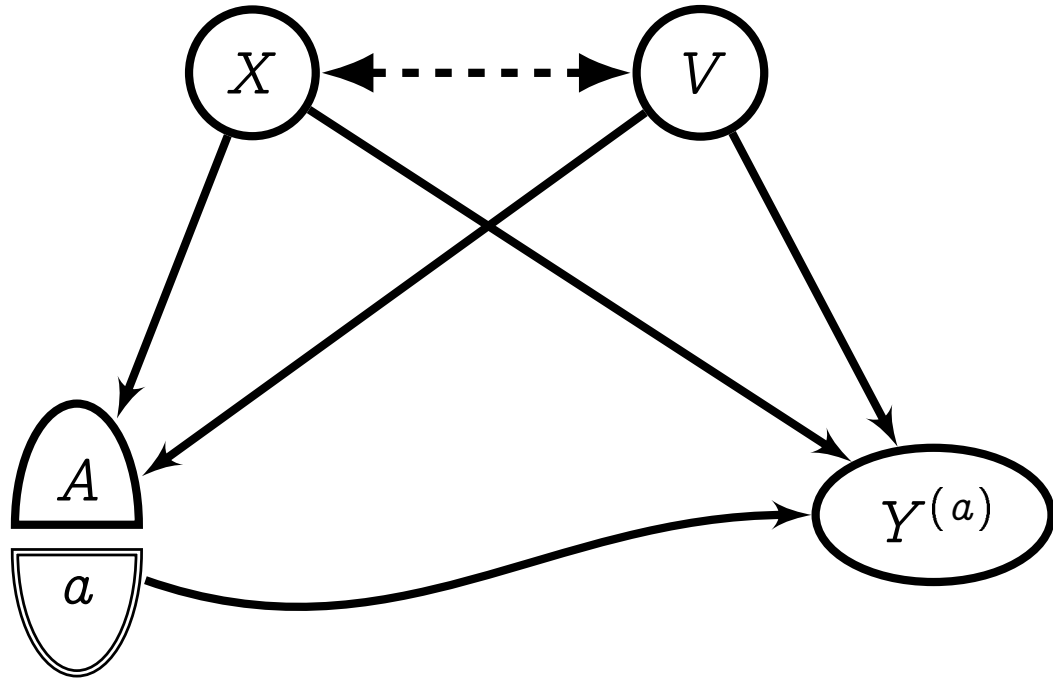
$$\theta_o^{\mathrm{CATE}}(a,v)$$
$$= \mathbb{E}(\,Y^{(a)} \,|\, V = v)$$
$$= \mathbb{E}_P\left(\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \,|\, V = v\right)$$



Conditional average treatment effect

Learned conditional mean:

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



Conditional ATE

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

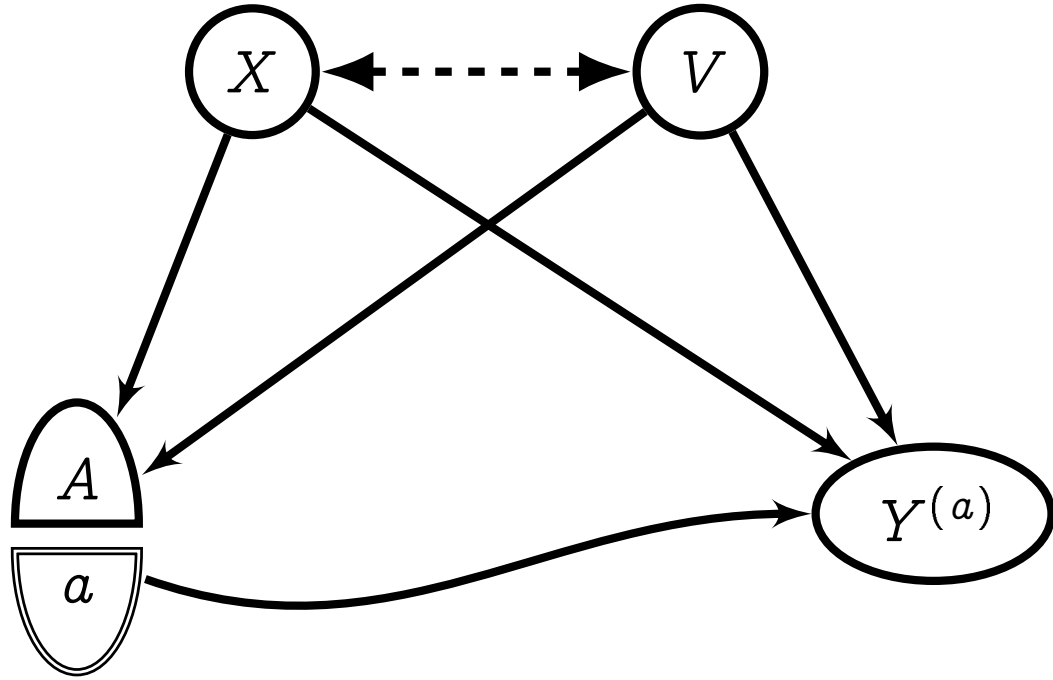
How to take conditional expectation?

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

Conditional average treatment effect

Learned conditional mean:

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



Conditional ATE

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

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 $E_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$E_0 \varphi(v) = \mu_{X|V=v}$$

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 $E_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$E_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$E_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff E_0 \in \text{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

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

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 $E_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$E_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$E_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff E_0 \in \text{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

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

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 $E_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$E_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$E_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff E_0 \in \text{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

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

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

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

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 $E_0 : \mathcal{H}_\mathcal{V} \to \mathcal{H}_\mathcal{X}$ such that

$$E_0 \varphi(v) = \mu_{X|V=v}$$

Assume

$$E_0 \in \overline{\text{span}\{\varphi(x) \otimes \varphi(v)\}} \iff E_0 \in \text{HS}(\mathcal{H}_\mathcal{V}, \mathcal{H}_\mathcal{X})$$

Implied smoothness assumption:

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

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

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

Ridge regression solution:

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

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

Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6]

$$E_0 = G_1 \circ T_1^{\frac{c_1-1}{2}}, \quad c_1 \in (1, 2], \quad \|G_1\|_{HS}^2 \leq \zeta_1,$$

T_1 is covariance of features $\varphi(v)$:

- Eigenspectrum decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

Larger $c_1 \implies$ smoother $E_0 \implies$ easier problem.

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

[B] Singh, Xu, G (2022a)

Earlier consistency proofs for finite dimensional $\varphi(x)$:

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).

Caponnetto, De Vito (2007).

Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6]

$$E_0 = G_1 \circ T_1^{\frac{c_1-1}{2}}, \quad c_1 \in (1,2], \quad \|G_1\|_{HS}^2 \leq \zeta_1,$$

T_1 is covariance of features $\varphi(v)$:

- Eigenspectrum decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

Larger $c_1 \implies$ smoother $E_0 \implies$ easier problem.

Consistency [A, Theorem 2, Theorem 3]

$$\left\| \widehat{E} - E_0 \right\|_{\text{HS}} = O_P\left(n^{-\frac{1}{2}\frac{c_1-1}{c_1+1/b_1}} \right),$$

best rate is $O_P(n^{-1/4})$ (minimax)

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

[B] Singh, Xu, G (2022a)

Earlier consistency proofs for finite dimensional $\varphi(x)$:

Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).

Caponnetto, De Vito (2007).

Consistency of CATE

Empirical CATE:

$$\hat{\theta}^{\mathrm{CATE}}(a, v)$$
$$= Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X \mid V=v}} \odot K_{Vv})$$

Consistency of CATE

Empirical CATE:

$$\hat{\theta}^{\mathrm{CATE}}(a, v) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot \underbrace{K_{XX}(K_{VV} + n\lambda_1 I)^{-1} K_{Vv}}_{\text{from } \hat{\mu}_{X \mid V=v}} \odot K_{Vv})$$

Consistency: [A, Theorem 2]

$$\|\hat{\theta}^{\mathrm{CATE}} - \theta_0^{\mathrm{CATE}}\|_\infty = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1//b}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1/b_1}}\right).$$

Follows from consistency of $\widehat{E}$ and $\hat{\gamma}$, under the assumptions:

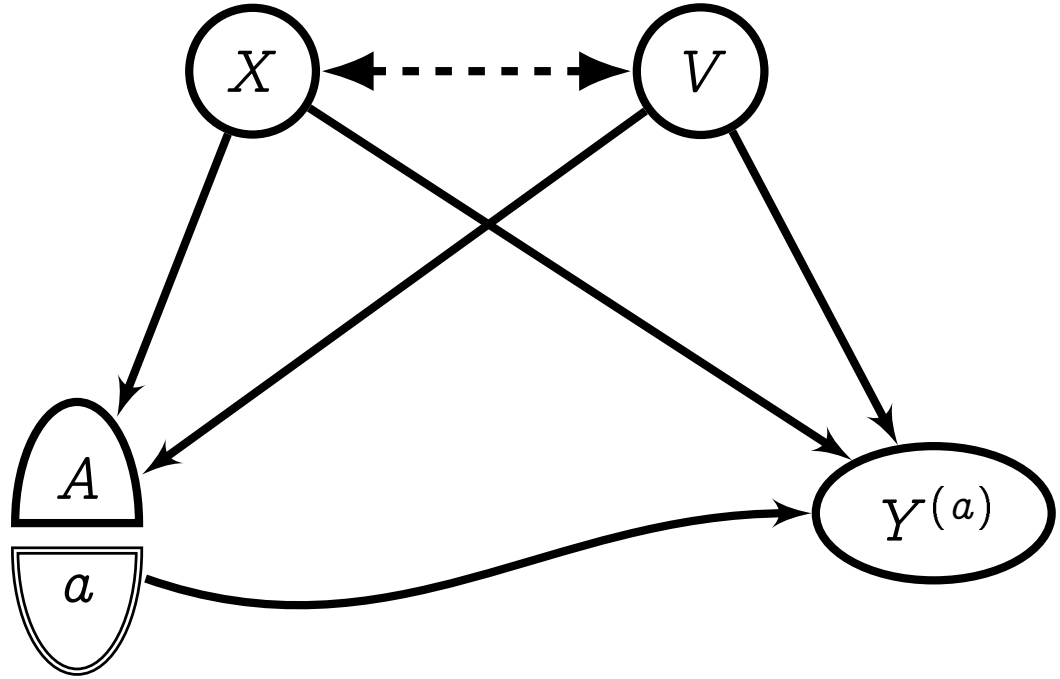
- $E_0 = G_1 \circ T_1^{\frac{c_1-1}{2}}$, $\|G_1\|_{HS}^2 \leq \zeta_1$,
- $\gamma_0 \in \mathcal{H}^c$.

[A] Singh, Xu, G (2022a)

Conditional ATE: example

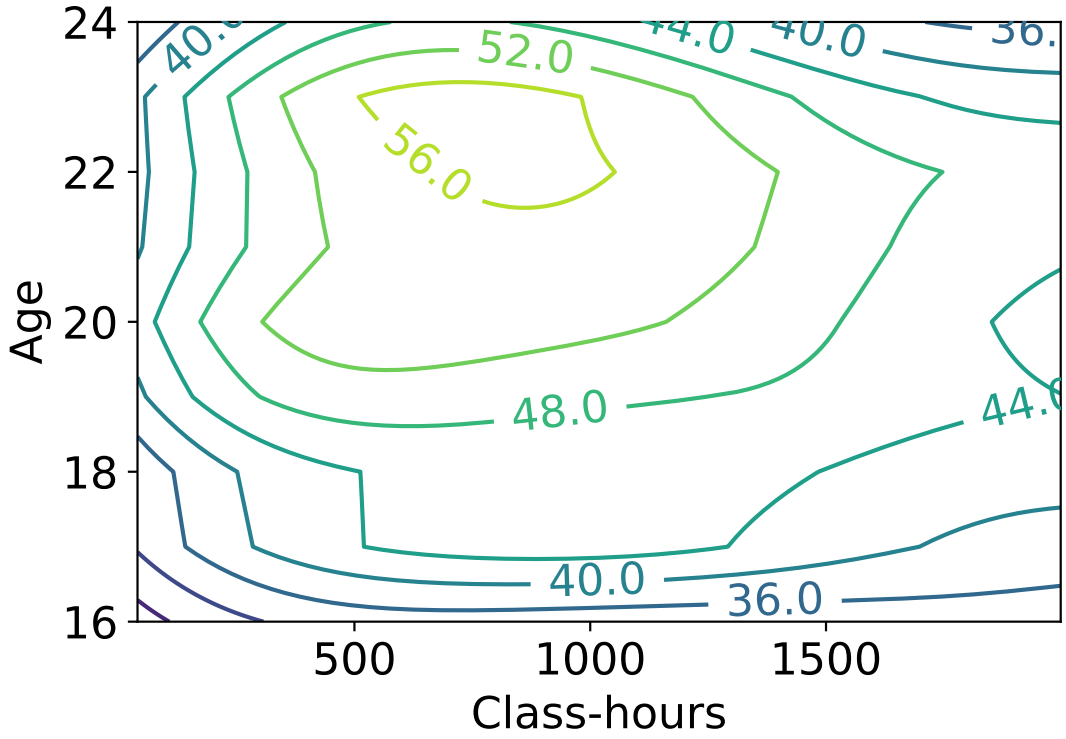
US job corps: training for disadvantaged youths:

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



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)

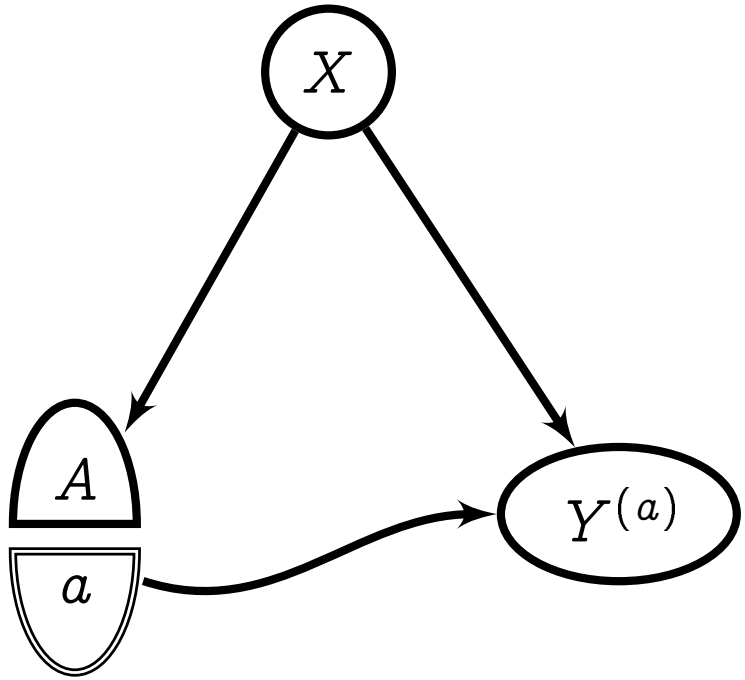
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\theta^{ATT}(a, a')$$
$$= \mathbb{E}(y^{(a')}|A = a)$$



Empirical ATT:

$$\hat{\theta}^{\mathrm{ATT}}(a, a')$$

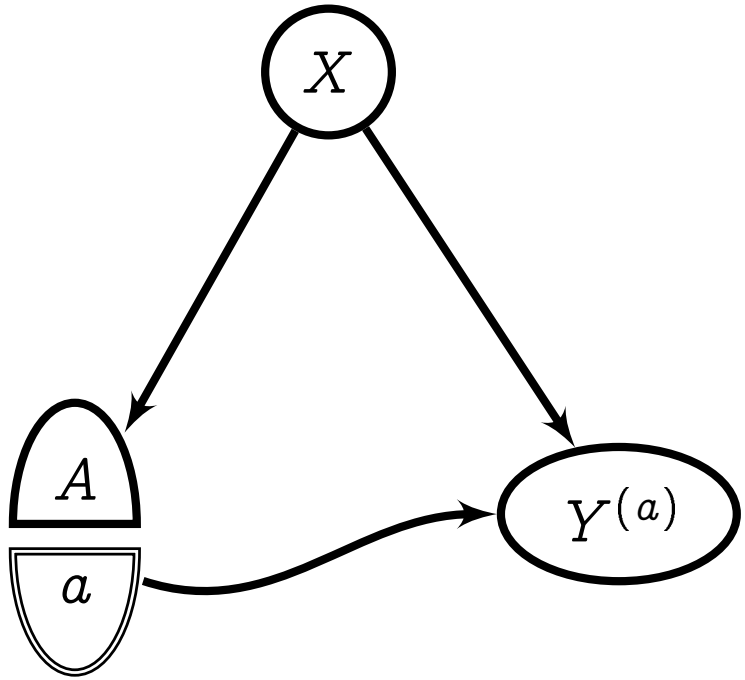
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\theta^{ATT}(a, a') = \mathbb{E}(y^{(a')}|A = a)$$



Empirical ATT:

$$\hat{\theta}^{\mathrm{ATT}}(a, a')$$

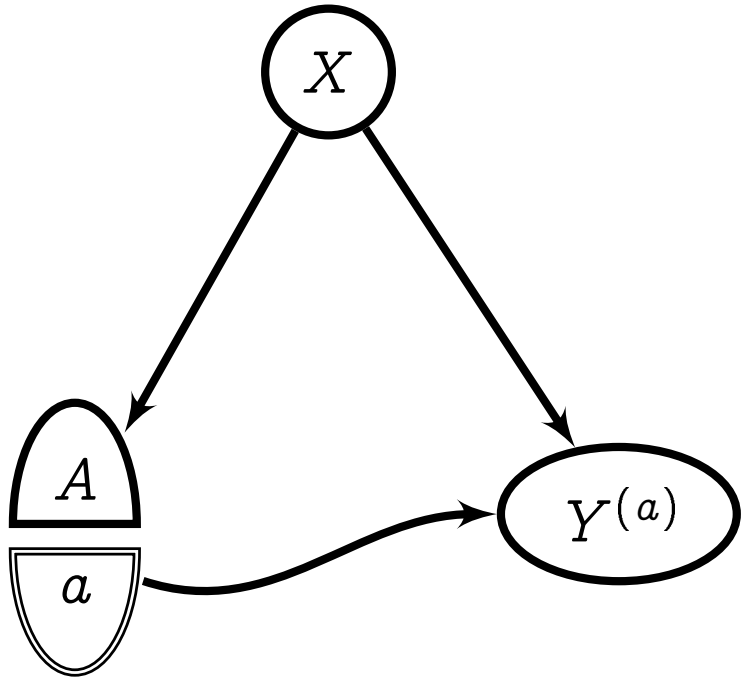
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\begin{aligned}
&\theta^{ATT}(a, a') \\
&= \mathbb{E}(y^{(a')}|A = a) \\
&= \mathbb{E}_P\left(\langle \gamma_0, \varphi(a') \otimes \varphi(X)\rangle \,|A = a\right) \\
&= \langle \gamma_0, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A = a]}_{\mu_{X|A=a}}\rangle
\end{aligned}$$



Empirical ATT:

$$\hat{\theta}^{\mathrm{ATT}}(a, a')$$

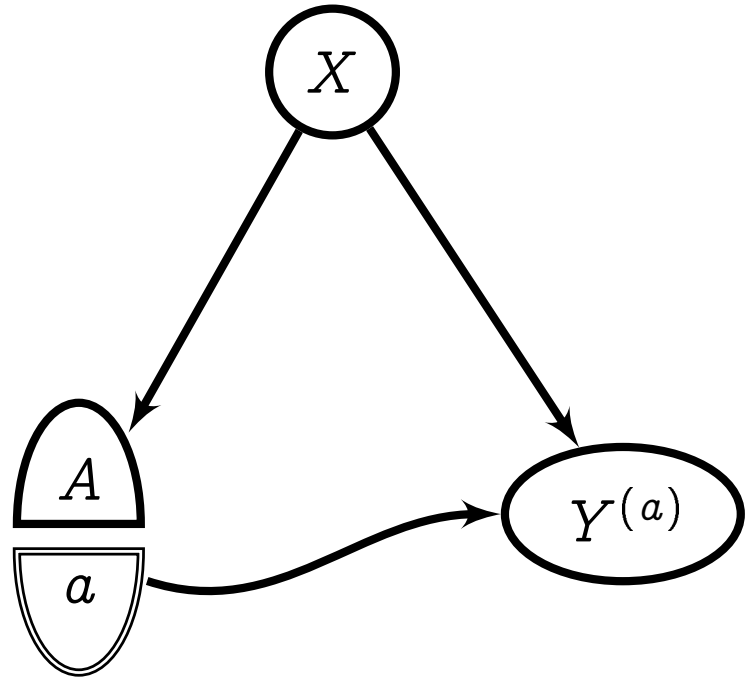
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\begin{aligned}
&\theta^{ATT}(a, a') \\
&= \mathbb{E}(y^{(a')}|A = a) \\
&= \mathbb{E}_P\left(\langle \gamma_0, \varphi(a') \otimes \varphi(X)\rangle\,|A = a\right) \\
&= \langle \gamma_0, \varphi(a') \otimes \underbrace{\mathbb{E}_P[\varphi(X)|A = a]}_{\mu_{X|A=a}}\rangle
\end{aligned}$$



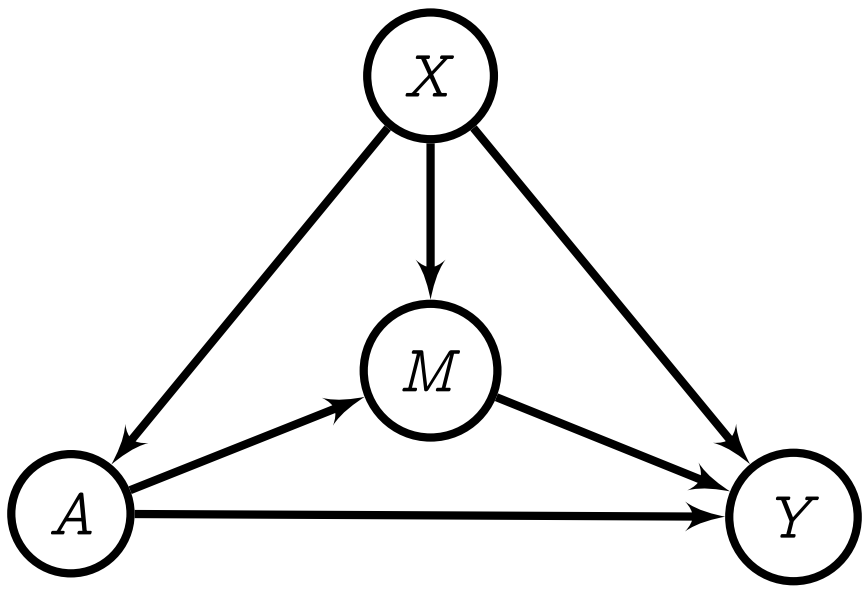
Empirical ATT:

$$\begin{aligned}
&\hat{\theta}^{\mathrm{ATT}}(a, a') \\
&= Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot \underbrace{K_{XX}(K_{AA} + n\lambda_1 I)^{-1} K_{Aa}}_{\text{from } \hat{\mu}_{X|A=a}})
\end{aligned}$$

Mediation analysis

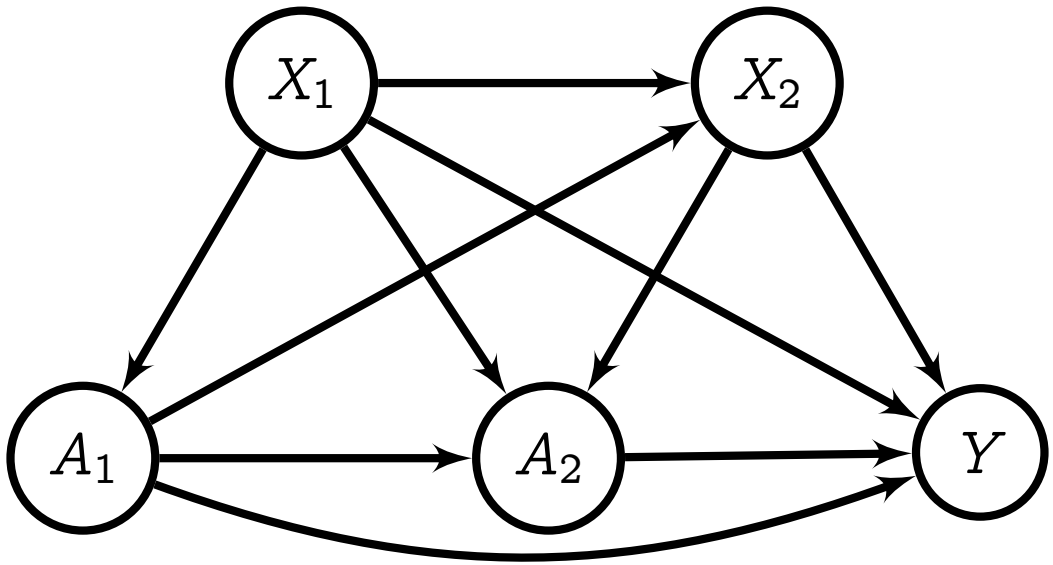
- Direct path from treatment A to effect Y
- Indirect path $A \rightarrow M \rightarrow Y$
- X: context

Is the effect Y mainly due to A? To M?



...dynamic treatment effect...

Dynamic treatment effect: sequence A_1, A_2 of treatments.

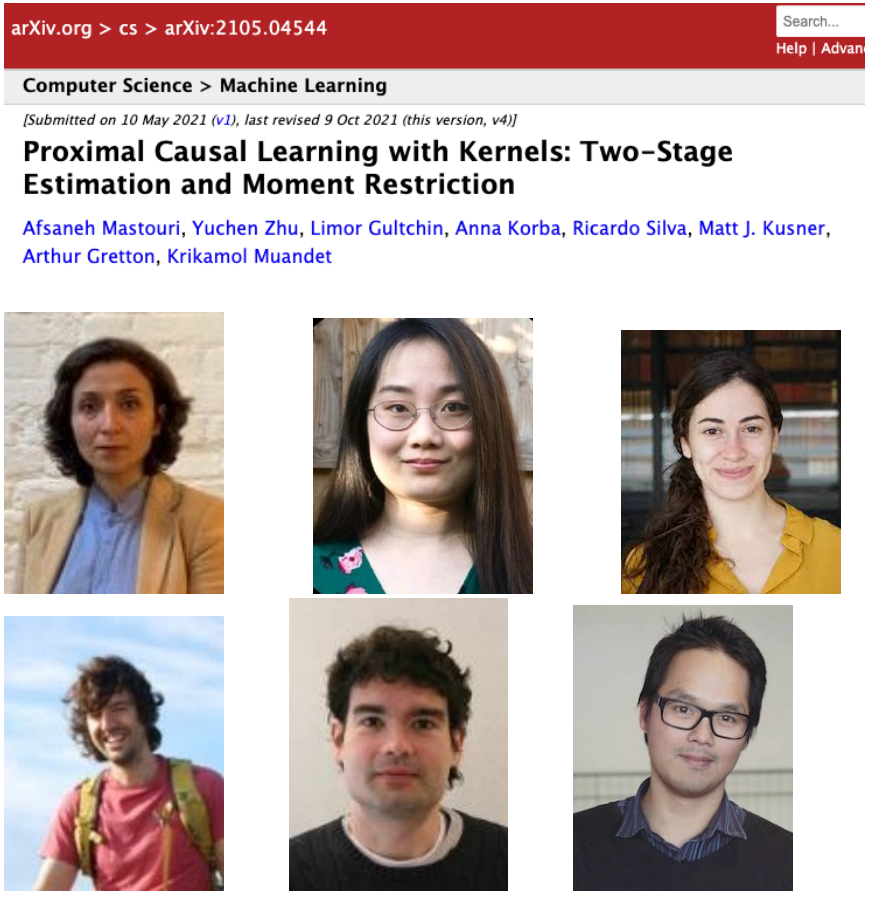


- potential outcomes $Y^{(a_1)}$, $Y^{(a_2)}$, $Y^{(a_1,a_2)}$,
- counterfactuals $\mathbb{E}(y^{(a_1',a_2')}|A_1 = a_1, A_2 = a_2)$...

(c.f. the Robins G-formula)

Unobserved confounders: proxy methods

Kernel features (ICML 2021):



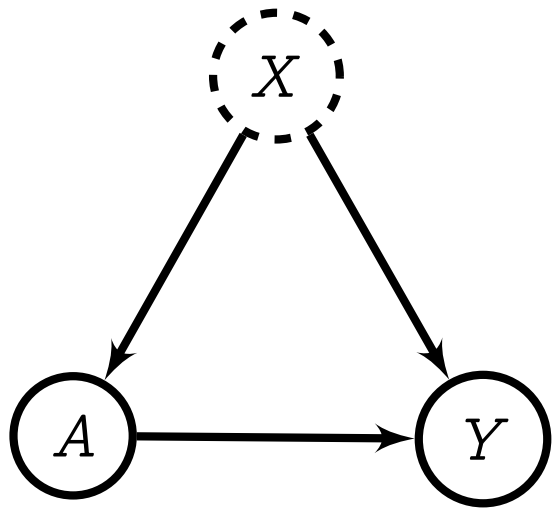
NN features (NeurIPS 2021):

The proxy correction

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

The definitions are:

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



If X were observed (which it isn't),

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

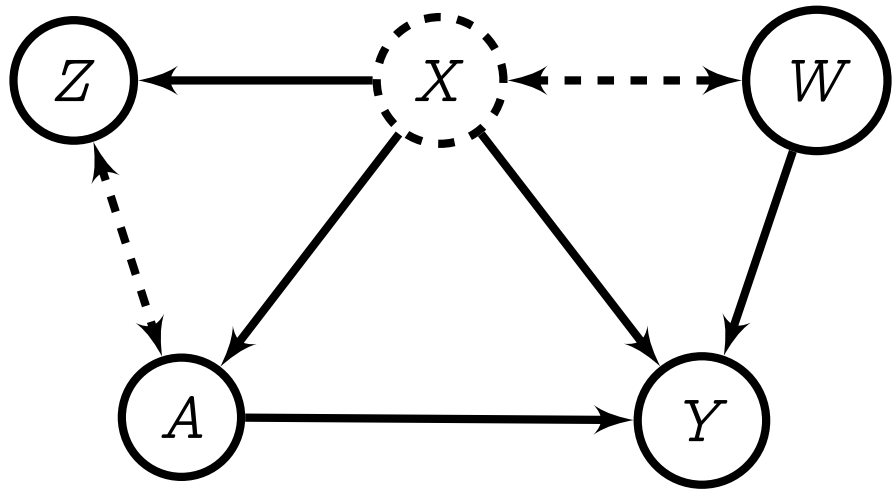
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder..

The proxy correction

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



Bidirected arrow: possible confounding.

Structural assumption:

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

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

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

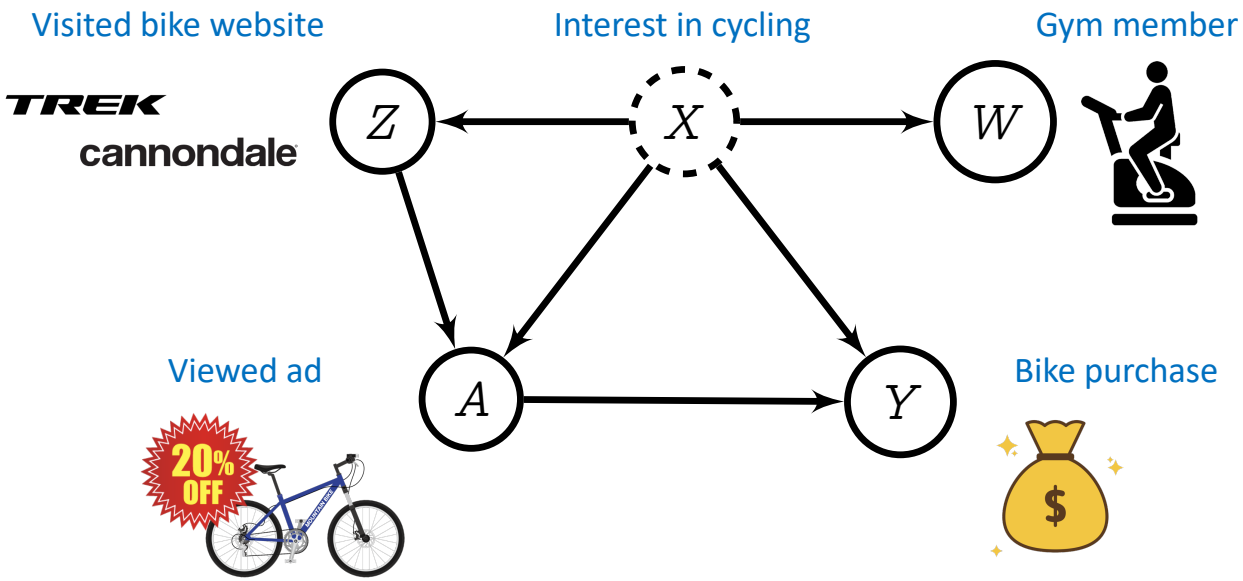
Miao, Geng, Tchetgen Tchetgen (2018): Identifying causal effects with proxy variables of an unmeasured confounder..

The proxy correction

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..

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

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

The proxy correction

If X were observed,

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

....but we do not see $p(x)$.

Miao, Geng, Tchetgen Tchetgen (2018)
Deaner (2021) Proxy controls and panel data.

The proxy correction

If X were observed,

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

....but we do not see $p(x)$.

Main theorem: Assume we have solved for bridge h_y...

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

(Fredholm integral of the first kind; subject to conditions for existence of solution)

Miao, Geng, Tchetgen Tchetgen (2018)
Deaner (2021) Proxy controls and panel data.

The proxy correction

If X were observed,

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

....but we do not see $p(x)$.

Main theorem: Assume we have solved for bridge h_y...

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

(Fredholm integral of the first kind; subject to conditions for existence of solution)

...then average causal effect via $p(w)$:

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

Expressions in terms of observed quantities, can be learned from data.

Miao, Geng, Tchetgen Tchetgen (2018)
Deaner (2021) Proxy controls and panel data.

Causal representation learning for proxies (1)

Bridge equation (previous slide):

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

Xu, Kanagawa, G. (2021) Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Causal representation learning for proxies (1)

Bridge equation (previous slide):

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

Squared loss for bridge ("stage 2"):

$$\mathcal{L}_2(h) = \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A,W)|A,Z])^2$$

Xu, Kanagawa, G. (2021) Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Causal representation learning for proxies (1)

Bridge equation (previous slide):

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

Squared loss for bridge ("stage 2"):

$$\mathcal{L}_2(h) = \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A,W)|A,Z])^2$$

Assume NN features ϕ_{θ_w} with weights θ_w, and bridge of form

$$h(a,w) = h^\top(\phi_{\theta_a}(a) \otimes \phi_{\theta_w}(w)).$$

Xu, Kanagawa, G. (2021) Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Causal representation learning for proxies (1)

Bridge equation (previous slide):

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

Squared loss for bridge ("stage 2"):

$$\mathcal{L}_2(h) = \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A,W)|A,Z])^2$$

Assume NN features ϕ_{θ_w} with weights θ_w, and bridge of form

$$h(a,w) = h^\top(\phi_{\theta_a}(a) \otimes \phi_{\theta_w}(w)).$$

Then

$$\mathbb{E}[h(A,W)|a,z] = h^\top\Big(\phi_{\theta_a}(a) \otimes \underbrace{\mathbb{E}[\phi_{\theta_w}(W)|a,z]}_{\mu_{W|a,z}}\Big)$$

$\mu_{W|a,z}$ is neural conditional mean embedding.

Xu, Kanagawa, G. (2021) Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Causal representation learning for proxies (2)

Our challenges:

1. How to obtain $\mu_{W|a,z} := \mathbb{E}_W[\phi_{\theta_w}(W)|a,z]$ for fixed θ_w?
2. How to optimize θ_w?

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (2)

Our challenges:

1. How to obtain $\mu_{W|a,z} := \mathbb{E}_W [\phi_{\theta_w}(W)|a,z]$ for fixed θ_w?
2. How to optimize θ_w?

Challenge 1: neural conditional mean embedding $\mu_{W|a,z}$ by ridge regression ("Stage 1"):

$$\hat{E}_{\theta_w} = \underset{E}{\operatorname{argmin}}\, \mathbb{E}_{WAZ} \|\phi_{\theta_w}(W) - E\phi_\gamma(A,Z)\|^2 + \lambda_1 \|E\|^2_{HS}$$

$$\mu_{W|a,z} = \hat{E}_{\theta_w}\phi_\gamma(a,z)$$

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (2)

Our challenges:

1. How to obtain $\mu_{W|a,z} := \mathbb{E}_W[\phi_{\theta_w}(W)|a,z]$ for fixed θ_w?
2. How to optimize θ_w?

Challenge 1: neural conditional mean embedding $\mu_{W|a,z}$ by ridge regression ("Stage 1"):

$$\hat{E}_{\theta_w} = \underset{E}{\operatorname{argmin}}\, \mathbb{E}_{WAZ}\|\phi_{\theta_w}(W) - E\phi_\gamma(A,Z)\|^2 + \lambda_1\|E\|^2_{HS}$$

$$\mu_{W|a,z} = \hat{E}_{\theta_w}\phi_\gamma(a,z)$$

$\hat{E}_{\theta_w}$ in closed form wrt $\phi_{\theta_w}, \phi_\gamma$: plug it in, take gradient steps for γ

(...but not θ_w - why not?)

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (3)

Challenge 2: optimize θ_w by plugging in the Stage 1 solution!

$$\mathcal{L}_2(h) = \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A, W)|A, Z])^2$$

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (3)

Challenge 2: optimize θ_w by plugging in the Stage 1 solution!

$$\mathcal{L}_2(h) = \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A, W)|A, Z])^2$$
$$= \mathbb{E}_{YAZ}\left[Y - h^\top \left(\phi_{\theta_a}(A) \otimes \mu_{W|A,Z}\right)\right]^2$$

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (3)

Challenge 2: optimize θ_w by plugging in the Stage 1 solution!

$$\begin{aligned}\mathcal{L}_2(h) &= \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A, W)|A, Z])^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \mu_{W|A,Z}\right)\right]^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \left(\hat{E}_{\theta_w}\phi_\gamma(A, Z)\right)\right)\right]^2\end{aligned}$$

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (3)

Challenge 2: optimize θ_w by plugging in the Stage 1 solution!

$$\begin{aligned}\mathcal{L}_2(h) &= \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A, W)|A, Z])^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \mu_{W|A,Z}\right)\right]^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \left(\hat{E}_{\theta_w}\phi_\gamma(A, Z)\right)\right)\right]^2\end{aligned}$$

$\hat{h}_y$ in closed form wrt $\phi_{\theta_w}, \phi_{\theta_a}$ by ridge regression:

$$\hat{h}_y := \underset{h}{\operatorname{argmin}}\, \mathcal{L}_2(h) + \lambda_1 \|h\|^2.$$

Xu, Kanagawa, G. (2021)

Causal representation learning for proxies (3)

Challenge 2: optimize θ_w by plugging in the Stage 1 solution!

$$\begin{aligned}\mathcal{L}_2(h) &= \mathbb{E}_{YAZ}(Y - \mathbb{E}[h(A,W)|A,Z])^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \mu_{W|A,Z}\right)\right]^2 \\ &= \mathbb{E}_{YAZ}\left[Y - h^\top\left(\phi_{\theta_a}(A) \otimes \left(\hat{E}_{\theta_w}\phi_\gamma(A,Z)\right)\right)\right]^2\end{aligned}$$

$\hat{h}_y$ in closed form wrt $\phi_{\theta_w}, \phi_{\theta_a}$ by ridge regression:

$$\hat{h}_y := \operatorname*{argmin}_h \mathcal{L}_2(h) + \lambda_1 \|h\|^2.$$

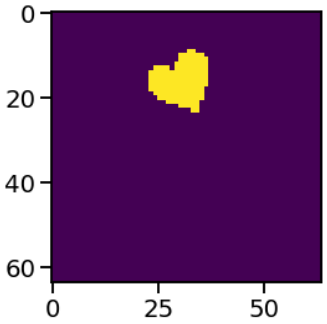
Plug in $\hat{h}_y$, take gradient steps on θ_a, θ_w

....but γ changes with θ_w

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

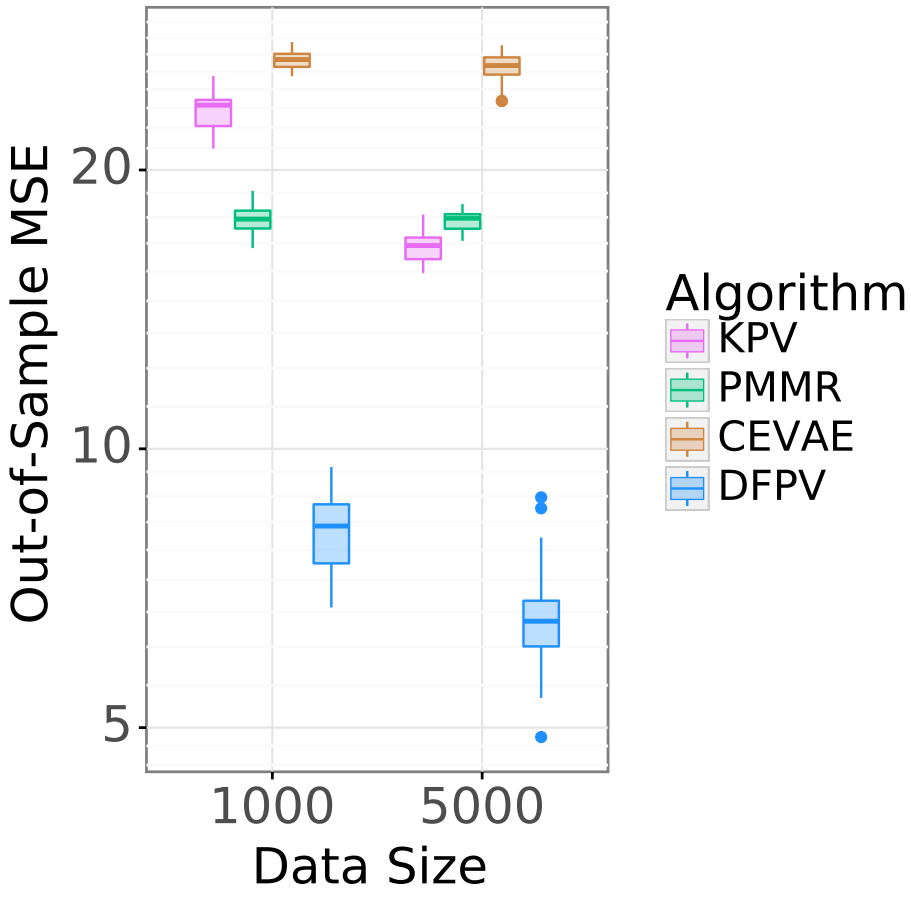
Xu, Kanagawa, G. (2021)

Synthetic experiment



Dsprite example:

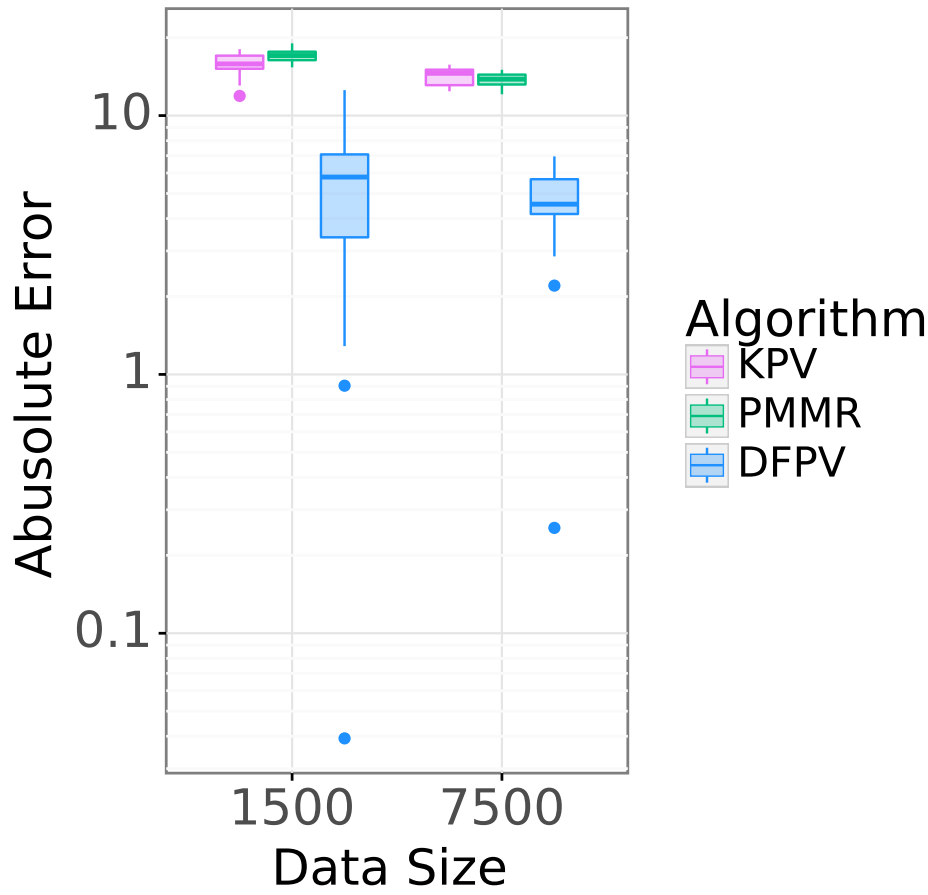
- $X =$ $\{\texttt{scale}, \texttt{rotation}, \texttt{posX}, \texttt{posY}\}$
- Treatment A is the image generated (with Gaussian noise)
- Outcome Y is quadratic function of A with multiplicative confounding by $\texttt{posY}$.
- $Z = \{\texttt{scale}, \texttt{rotation}, \texttt{posX}\}$, $W =$noisy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



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.



Conclusions

Neural net and kernel solutions:

- ...for ATE, ATT, CATE, mediation analysis, dynamic treatment effects
- ...even for unobserved covariates (proxy methods)
- ...with treatment A, covariates X, V, proxies (W, Z) multivariate, "complicated"
- Convergence guarantees for kernels and NN

Not in this talk:

- Elasticities
- Regression to potential outcome distributions over Y (not just $E(Y^{(a)}|\ldots)$)
- Instrumental variable regression (kernel and NN)

Code available for NN and kernel proxy methods:

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

Selected papers

Observed confounders:

ICLR 2023:

Unobserved confounders:

ICML 2021:

NeurIPS 2021:

NeurIPS 2019:

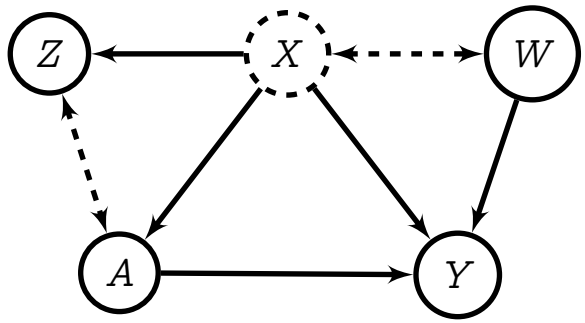
Questions?


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

Proxy proof (discrete variables)

If X were observed,

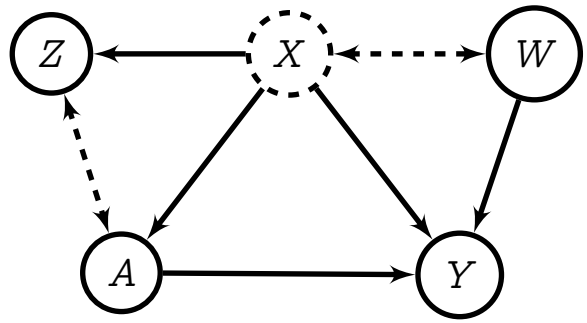
$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i)$$



Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$



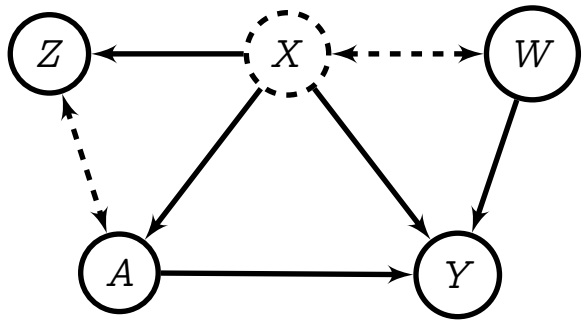
Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$

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

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



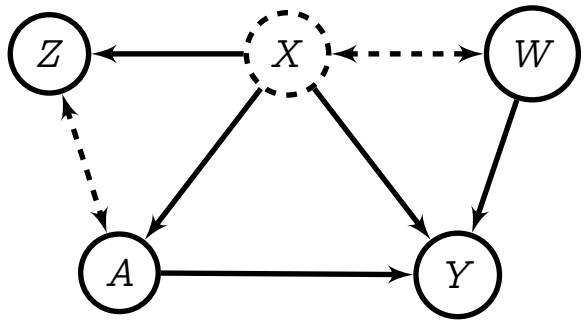
Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$

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

$$P(W|Z, a) = P(W|X)P(X|Z, a)$$
$$\implies P(X|Z, a) = P^{-1}(W|X)P(W|Z, a)$$



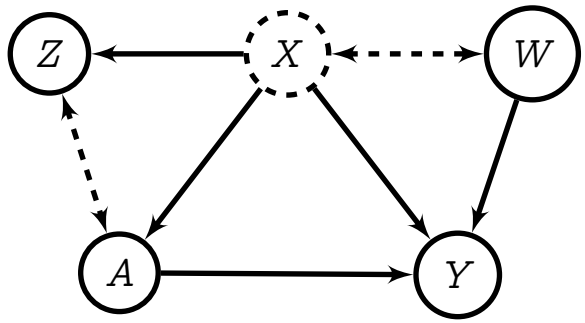
Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$

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

$$P(W|Z, a) = P(W|X)P(X|Z, a)$$
$$\implies P(X|Z, a) = P^{-1}(W|X)P(W|Z, a)$$



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

$$P(y|Z, a) = P(y|X, a)P(X|Z, a)$$

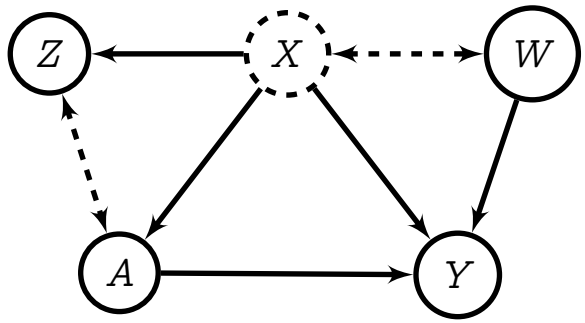
Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$

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

$$P(W|Z, a) = P(W|X)P(X|Z, a)$$
$$\implies P(X|Z, a) = P^{-1}(W|X)P(W|Z, a)$$



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

$$P(y|Z, a) = P(y|X, a)\underbrace{P^{-1}(W|X)P(W|Z, a)}_{P(X|Z,a)}$$

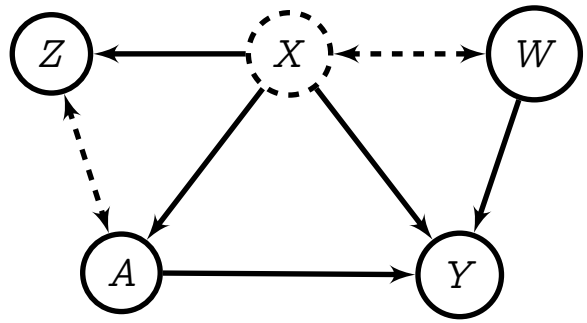
Proxy proof (discrete variables)

If X were observed,

$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i) = P(y|X, a)P(X)$$

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

$$P(W|Z, a) = P(W|X)P(X|Z, a)$$
$$\implies P(X|Z, a) = P^{-1}(W|X)P(W|Z, a)$$



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

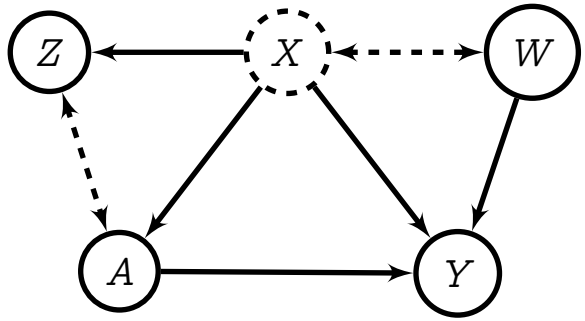
$$P(y|Z, a) = P(y|X, a)\underbrace{P^{-1}(W|X)P(W|Z, a)}_{P(X|Z,a)}$$

$$\implies p(y|X, a) = p(y|Z, a)P^{-1}(W|Z, a)P(W|X)$$

Proof (discrete variables)

From previous slide:

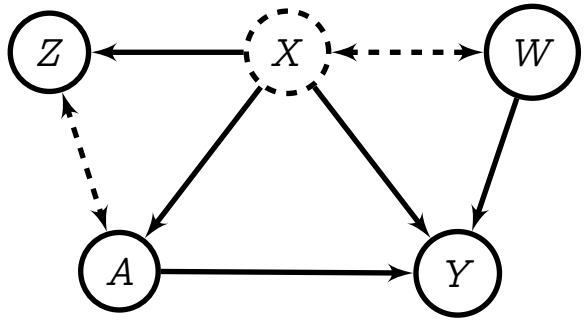
$$p(y|X,a) = p(y|Z,a)P^{-1}(W|Z,a)P(W|X)$$



Proof (discrete variables)

From previous slide:

$$p(y|X,a) = p(y|Z,a)P^{-1}(W|Z,a)P(W|X)$$



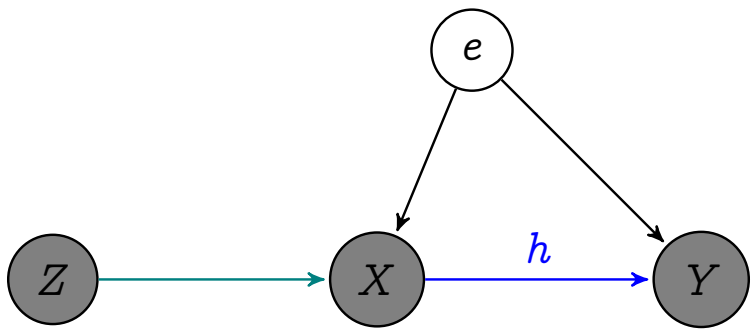
Multiply LHS and RHS by $P(X)$:

$$\begin{aligned} P(Y^{(a)}) &:= P(y|X,a)P(X) \\ &= p(y|Z,a)P^{-1}(W|Z,a)\underbrace{P(W|X)P(X)}_{P(W)} \end{aligned}$$

Average causal effect using only observed data!

Instrumental variable setting (1)

- Unobserved confounder $e \implies$ prediction $\neq$ counterfactual prediction
- goal: learn causal relationship h between input X and output Y
  - if we intervened on X, what would be the effect on Y?
- Instrument Z only influences Y via X, identifying h

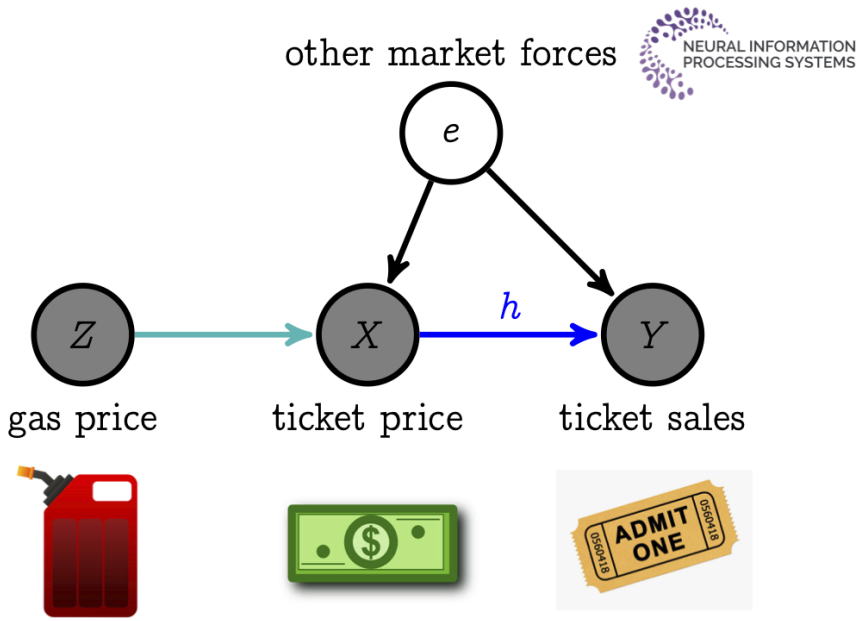


$$Y = \langle h, \psi(X) \rangle + e \qquad \mathbb{E}(e|Z) = 0$$

Singh, Sahani, G., (NeurIPS 2019)
Xu, Chen, Srinivasan, de Freitas, Doucet, G. (ICLR 2021)

Instrumental variable setting (1)

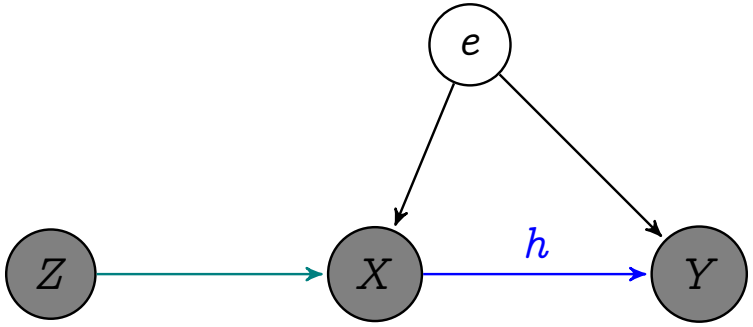
- Unobserved confounder $e \implies$ prediction $\neq$ counterfactual prediction
- goal: learn causal relationship h between input X and output Y
  - if we intervened on X, what would be the effect on Y?
- Instrument Z only influences Y via X, identifying h



$$Y = \langle h, \psi(X) \rangle + e \qquad \mathbb{E}(e|Z) = 0$$

Singh, Sahani, G., (NeurIPS 2019)
Xu, Chen, Srinivasan, de Freitas, Doucet, G. (ICLR 2021)

Instrumental variable setting (2)



- Ridge regression of $\psi(X)$ on $\phi(Z)$
  - using n observations
  - construct conditional mean embedding $\mu(z) := \mathbb{E}[\psi(X)|Z = z]$
- Ridge regression of Y on $\mu(Z)$
  - using remaining m observations
  - this is the estimator for h
- Solved using kernel and learned NN features

Singh, Sahani, G., (NeurIPS 2019)
Xu, Chen, Srinivasan, de Freitas, Doucet, G. (ICLR 2021)