Causal modelling with kernels: treatment effects, counterfactuals, mediation, and proxies

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

A medical treatment scenario

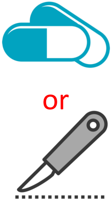
or

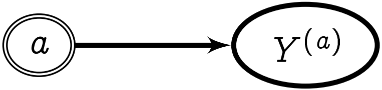
From our observations of historical hospital data:

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

Just recommend pills? Cheaper and more effective!

A medical treatment scenario





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

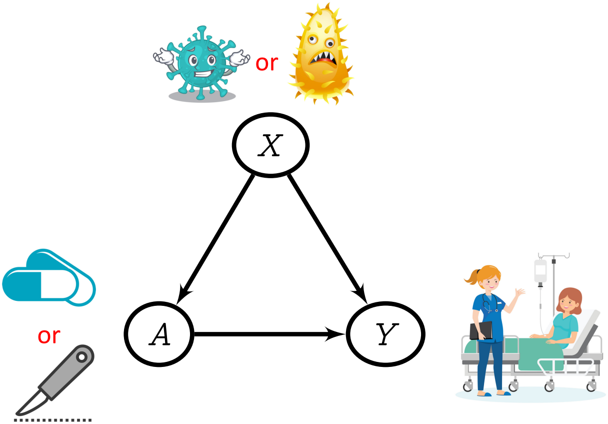
- $P(Y = \text{cured} | do(\text{pills})) = 0.64$
- $P(Y = \text{cured} | do(\text{surgery})) = 0.75$

What went wrong?

Observational vs interventional

Conditioning from observation:

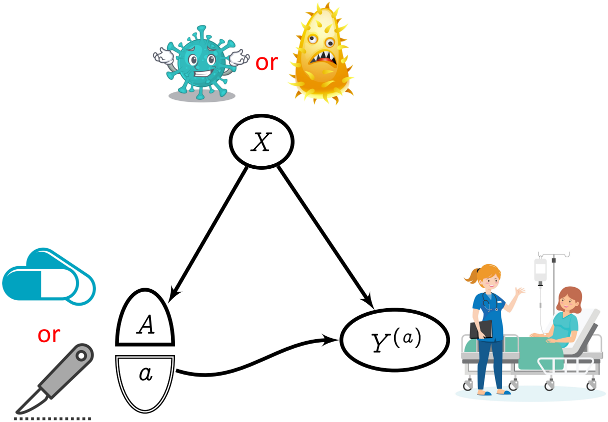
$$E(\,Y\,|\,A = a) = \sum_{x} E(y|a, x)p(x|a)$$



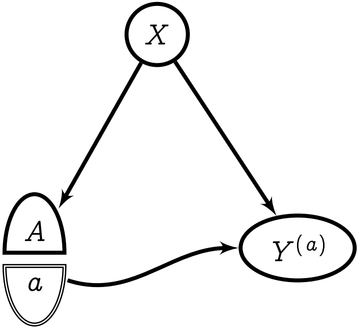
Observational vs interventional

Average causal effect (intervention):

$$\mathrm{E}(Y^{(a)}) = \sum_x E(y|a, x)p(x)$$



Questions we will solve



Outline

Talk structure:

- Average treatment effect (ATE)
  - ...via kernel/NN mean embedding (marginalization)
- Conditional average treatment effect (CATE)
  - via conditional mean embedding
- Average treatment on treated
- Mediation effect, dynamic treatment effect
- Proxy methods
  - ...when covariates are hidden

Properties and advantages of approach:

- Treatment A, covariates X, etc are by default multivariate, complicated...
- Simple, robust implementation;
- Strong statistical guarantees under general smoothness assumptions

Methods also implemented for 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

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)

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.

Primary focus of this talk.

Building block: kernel ridge regression

Learn $\gamma_0(x) := \mathrm{E}[Y|X=x]$ 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:

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

Building block: kernel ridge regression

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

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

Kernel as feature dot product:

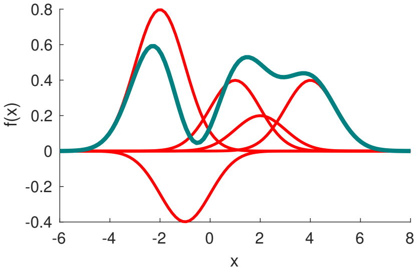
$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Solution at x:

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

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

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



Building block: kernel ridge regression

Learn $\gamma_0(x) := \mathrm{E}[Y|X=x]$ 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:

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

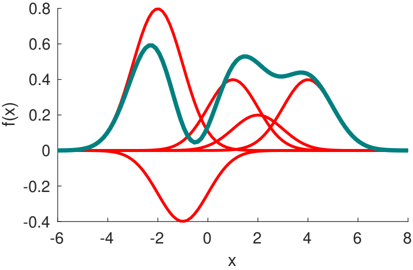
Solution at x (as weighted sum of y)

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

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

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

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



KRR: consistency in RKHS norm

Assume problem well specified

- Denote: $\gamma_0 \in \mathcal{H}^c$ where $\mathcal{H}^c \subset \mathcal{H}, \quad c \in (1,2]$
- Larger $c \implies$ smoother $\gamma_0 \implies$ easier problem.

[A] Singh, Xu, G (2021a), Generalized Kernel Ridge Regression for Nonparametric Structural Functions and Semiparametric Treatment Effects.

Results from:

Smale and Ding-Xuan Zhou (2007). Learning theory estimates via integral operators and their approximations; Caponnetto, De Vito (2007), Optimal rates for the regularized least-squares algorithm.

KRR: consistency in RKHS norm

Assume problem well specified

- Denote: $\gamma_0 \in \mathcal{H}^c$ where $\mathcal{H}^c \subset \mathcal{H}, \quad c \in (1, 2]$
- Larger $c \implies$ smoother $\gamma_0 \implies$ easier problem.

Consistency [A, Prop. F.1]

$$\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right),$$

best rate is $O_P(n^{-1/6})$.

[A] Singh, Xu, G (2021a), Generalized Kernel Ridge Regression for Nonparametric Structural Functions and Semiparametric Treatment Effects.

Results from:

Smale and Ding-Xuan Zhou (2007). Learning theory estimates via integral operators and their approximations; Caponnetto, De Vito (2007), Optimal rates for the regularized least-squares algorithm.

(Conditional) average treatment effect, average treatment on treated

Average treatment effect

Average causal effect (intervention):

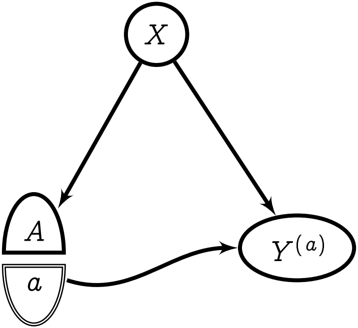
$$\mathrm{E}(Y^{(a)}) = \int E(y|a,x)dp(x)$$

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

Assume: (1) no interference/spillover, (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, ...)



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

Multiple inputs via products of kernels

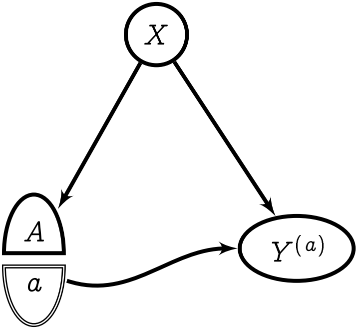
We may predict expected outcome from two inputs

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

Assume we have:

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

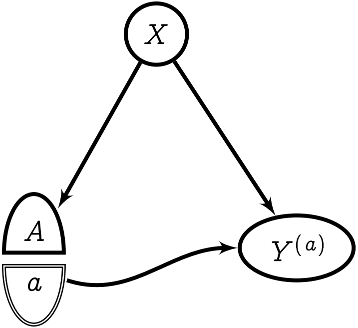
(argument of kernel/feature map indicates feature space)



Multiple inputs via products of kernels

We may predict expected outcome from two inputs

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



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)

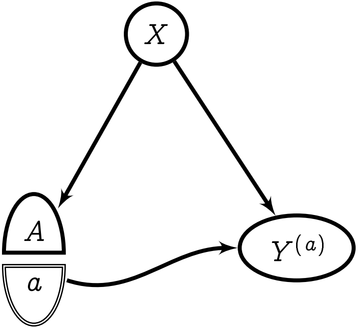
We use outer product of features ($\Longrightarrow$ product of kernels):

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

Multiple inputs via products of kernels

We may predict expected outcome from two inputs

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



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:

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

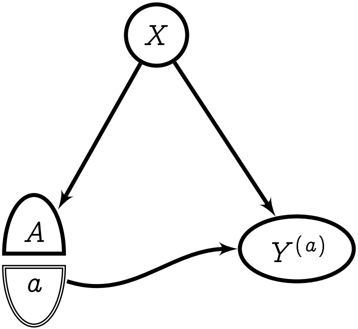
ATE (dose-response curve)

Well specified setting:

$$\gamma_0(a, x) = \mathrm{E}[\,Y\,|\,a, x] \in \mathcal{H}$$

ATE as feature space dot product:

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



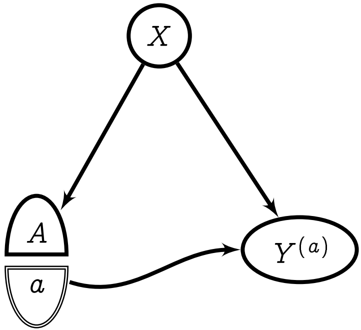
ATE (dose-response curve)

Well specified setting:

$$\gamma_0(a, x) = \mathrm{E}[Y|a, x] \in \mathcal{H}$$

ATE as feature space dot product:

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



Feature map of probability P,

$$\mu_P = [\ldots \mathrm{E}_P[\varphi_i(X)] \ldots]$$

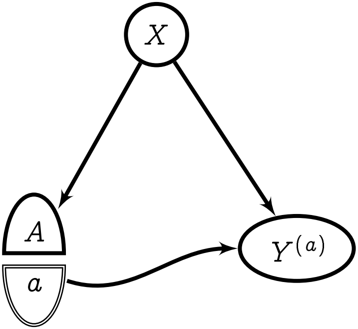
ATE (dose-response curve)

Well specified setting:

$$\gamma_0(a, x) = \mathrm{E}[Y|a, x] \in \mathcal{H}$$

ATE as feature space dot product:

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



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:

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

ATE: empirical estimate and consistency

Empirical estimate of ATE:

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

Consistency:

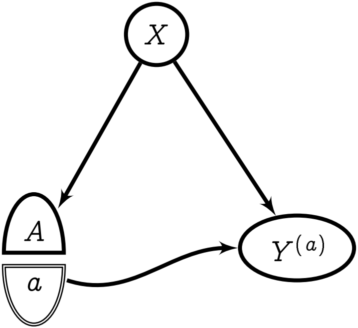
$$\left\|\hat{\theta}^{\mathrm{ATE}} - \theta_o^{\mathrm{ATE}}\right\|_\infty = O_P\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right)$$

Follows from consistency of $\hat{\mu}_P$, and of $\hat{\gamma}$ under smoothness assumption $\gamma_0 \in \mathcal{H}^c$.

ATE: example

US job corps: training for disadvantaged youths:

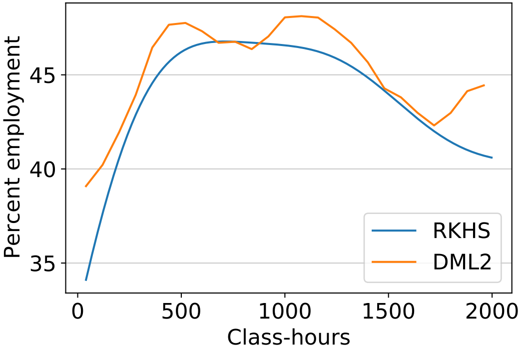
- X: covariate/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (percent employment)



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

Singh, Xu, G (2021a).

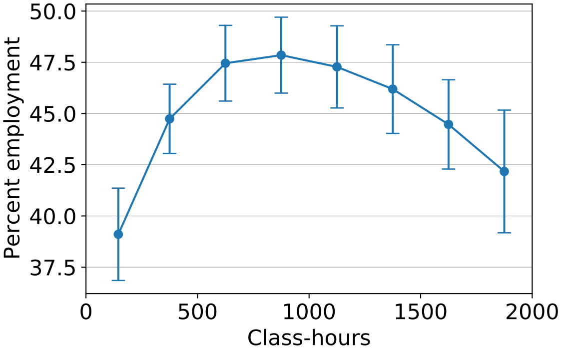
ATE: results



- 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 (2021a)

Confidence intervals for discretized treatment



- Doubly robust estimator: semiparametric efficiency, asymptotic normality, confidence intervals
- Automated debiasing (via kernel regression)
- Requires discretized treatment (here, equiprobable bins)

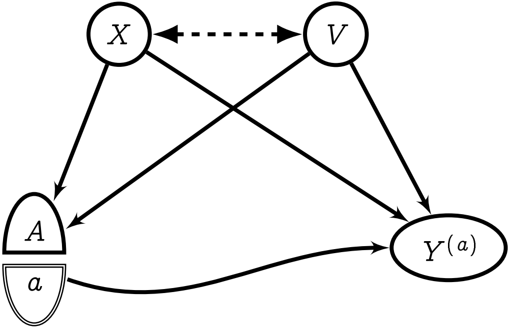
Singh, Xu, G (2021a)

Chernozhukov, Newey, Singh (2018). Automatic debiased machine learning of causal and structural effects.

Conditional ATE: example

US job corps: training for disadvantaged youths:

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



Singh, Xu, G (2021a)

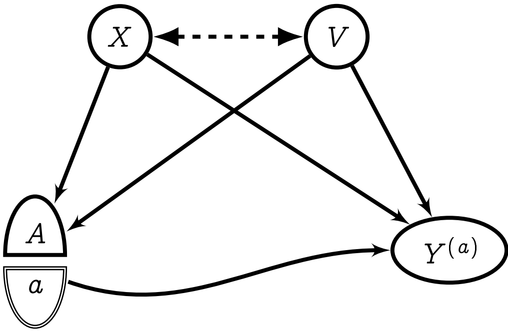
Conditional average treatment effect

Learned conditional mean:

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

Conditional ATE

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



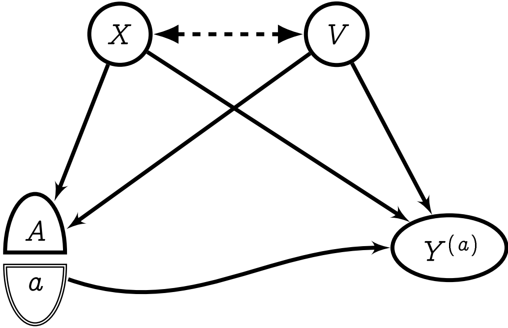
Conditional average treatment effect

Learned conditional mean:

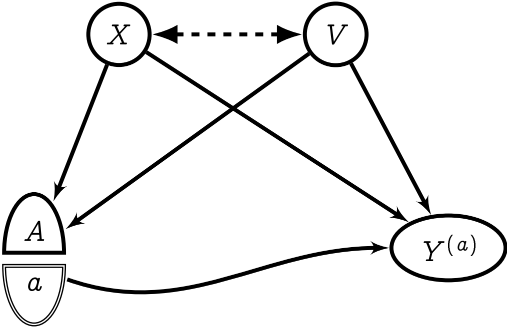
$$\mathrm{E}[Y|a,x,v] \approx \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a)\otimes\varphi(x)\otimes\varphi(v)\rangle\,.$$

Conditional ATE

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



Conditional average treatment effect



Learned conditional mean:

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

Conditional ATE

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

How to take conditional expectation?

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

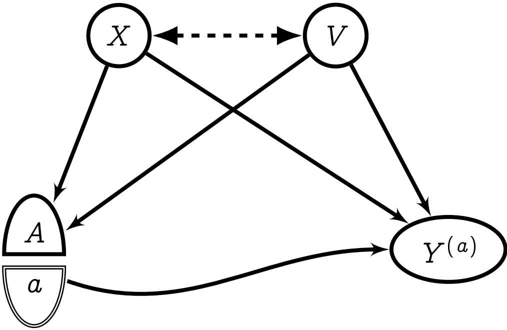
Conditional average treatment effect

Learned conditional mean:

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

Conditional ATE

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



Learn conditional mean embedding: $\mu_{X|V=v} := \mathrm{E}_P\left(\varphi(X)|V=v\right)$

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.

Singh, Sahani, G (2019), Kernel Instrumental Variable Regression.

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

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

Smoothness assumption:

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

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.

Singh, Sahani, G (2019), Kernel Instrumental Variable Regression.

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

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

Smoothness assumption:

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

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.

Singh, Sahani, G (2019), Kernel Instrumental Variable Regression.

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

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

Smoothness assumption:

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

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.

Singh, Sahani, G (2019), Kernel Instrumental Variable Regression.

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

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

Smoothness assumption:

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

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

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

Ridge regression solution:

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

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

Consistency of conditional mean embedding

Assume problem well specified [A, Hypothesis 5]

$$E_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}^{c_1}, \mathcal{H}_{\mathcal{X}})$$

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

[A] Singh, Sahani, G (2019)

Earlier consistency proof 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 [A, Hypothesis 5]

$$E_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}^{c_1}, \mathcal{H}_{\mathcal{X}})$$

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

Consistency [A, Theorem 2]

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

best rate is $O_P(n^{-1/6})$.

[A] Singh, Sahani, G (2019)

Earlier consistency proof for finite dimensional $\varphi(x)$:
Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).
Caponnetto, De Vito (2007).

Consistency of CATE

Empirical CATE:

$$\hat{\theta}^{\mathrm{CATE}}(a, v) = \left\langle \hat{\gamma}, \varphi(a) \otimes \hat{\mu}_{X \mid V=v} \otimes \varphi(v) \right\rangle$$

Consistency of CATE

Empirical CATE:

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

Consistency of CATE

Empirical CATE:

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

Consistency:

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

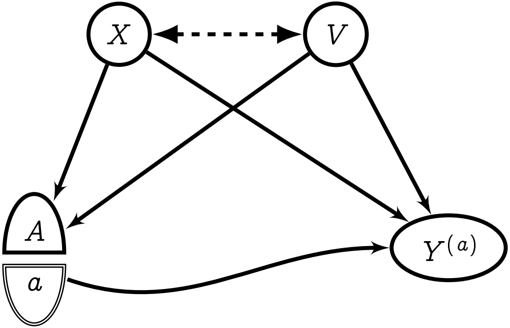
Follows from consistency of $\widehat{E}$ and $\hat{\gamma}$, under the smoothness assumptions.

Singh, Xu, G (2021a)

Conditional ATE: example

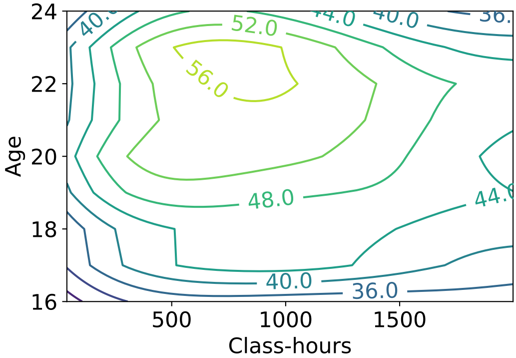
US job corps: training for disadvantaged youths:

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



Singh, Xu, G (2021a)

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

Singh, Xu, G (2021a)

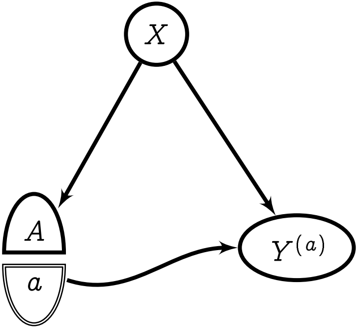
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

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



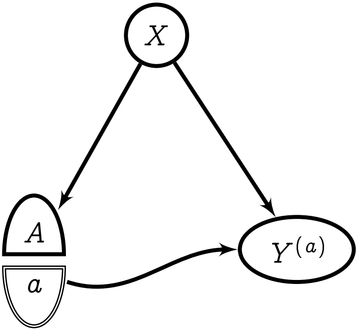
Empirical ATT:

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

Counterfactual: average treatment on treated

Conditional mean:

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



Average treatment on treated:

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

Empirical ATT:

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

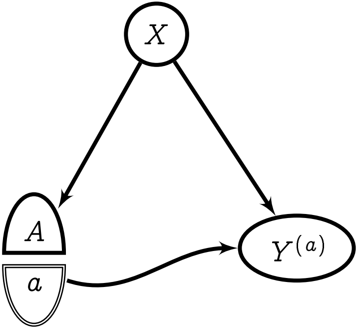
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\theta^{ATT}(a, a')$$
$$= \mathrm{E}(Y^{(a')}|A = a)$$
$$= \mathrm{E}_P\left(\langle\gamma_0, \varphi(a') \otimes \varphi(X)\rangle \,|A = a\right)$$
$$= \langle\gamma_0, \varphi(a') \otimes \underbrace{\mathrm{E}_P[\varphi(X)|A = a]}_{\mu_{X|A=a}}\rangle$$



Empirical ATT:

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

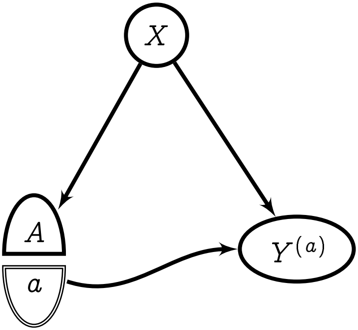
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

$$\theta^{ATT}(a,a') = \mathrm{E}(Y^{(a')}|A=a) = \mathrm{E}_P\left(\langle\gamma_0, \varphi(a')\otimes\varphi(X)\rangle\,|A=a\right) = \langle\gamma_0, \varphi(a')\otimes\underbrace{\mathrm{E}_P[\varphi(X)|A=a]}_{\mu_{X|A=a}}\rangle$$



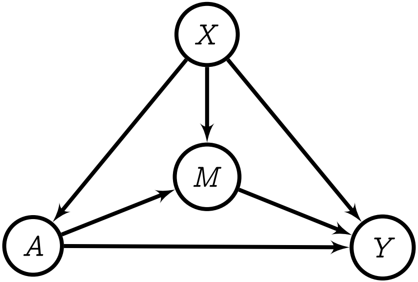
Empirical ATT:

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

Mediation analysis

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

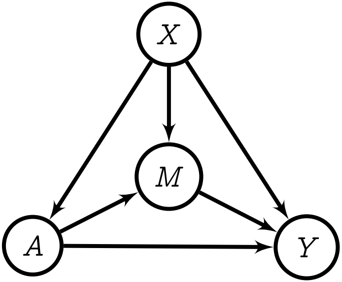
Is the effect Y mainly due to A? To M?



Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



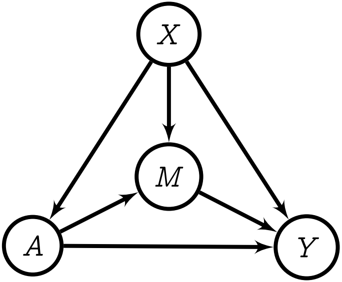
$\gamma_0(a, m, x) \approx \mathrm{E}[Y | A = a, M = m, X = x]$

Singh, Xu, G (2021b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$$\gamma_0(a, m, x) \approx \mathrm{E}[\, Y \,|\, A = a, M = m, X = x]$$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_0(a', M, X) \mathrm{d}\mathbb{P}(M | A = a, X) d\mathbb{P}(X)$$

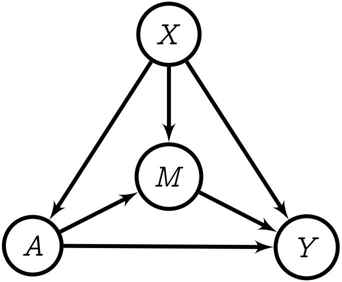
Effect of intervention a', with $M^{(a)}$ as if intervention were a

Singh, Xu, G (2021b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: example

US job corps: training for disadvantaged youths:

- X: confounder/context (age, education, marital status, ...)
- A: treatment (training hours)
- Y: outcome (arrests)
- M: mediator (employment)



$\gamma_0(a, m, x) \approx \mathrm{E}[\, Y \,|\, A = a, M = m, X = x]$

A quantity of interest, the mediated effect:

$$Y^{\{a', M^{(a)}\}} = \int \gamma_0(a', M, X) \mathrm{d}\mathbb{P}(M | A = a, X) d\mathbb{P}(X)$$
$$= \langle \gamma_0, \varphi(a') \otimes \mathrm{E}_P\{\mu_{M|A=a,X} \otimes \varphi(X)\}\rangle$$

Effect of intervention a', with $M^{(a)}$ as if intervention were a

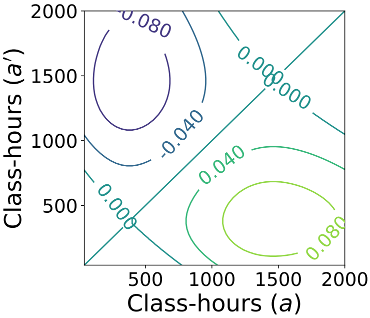
Singh, Xu, G (2021b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

Mediation analysis: results

Total effect:

$$\theta_0^{TE}(a, a')$$

$$:= \mathbb{E}[\, Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests

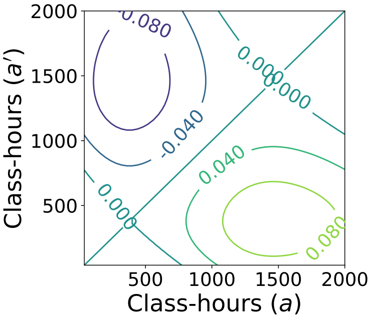
Singh, Xu, G (2021b)

Mediation analysis: results

Total effect:

$$\theta_0^{TE}(a, a')$$

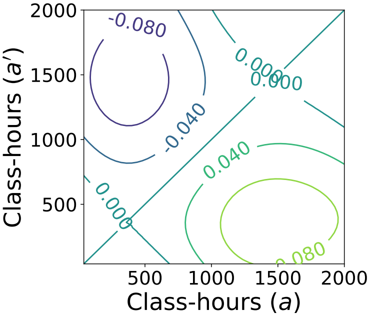
$$:= \mathbb{E}[\,Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



Direct effect:

$$\theta_0^{DE}(a, a')$$

$$:= \mathbb{E}[\,Y^{\{a', M^{(a)}\}} - Y^{\{a, M^{(a)}\}}]$$

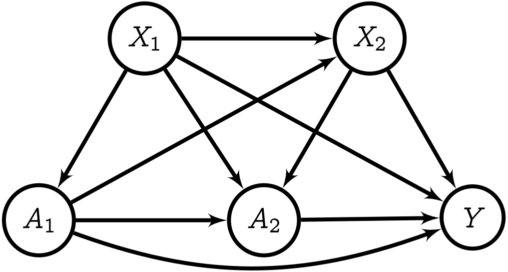


- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests
- Indirect effect mediated via employment effectively zero

Singh, Xu, G (2021b)

...dynamic treatment effect...

Dynamic treatment effect: sequence A_1, A_2 of treatments.



- Causal effects $Y^{(a_1)}, Y^{(a_2)}, Y^{(a_1,a_2)}$,
- counterfactuals $\mathrm{E}(y^{(a_1',a_2')} | A_1 = a_1, A_2 = a_2)$...

(c.f. the Robins G-formula)

Unobserved confounders

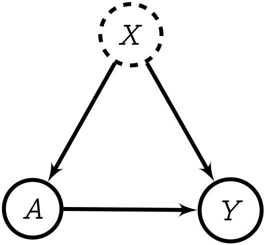
 

The proxy correction

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),

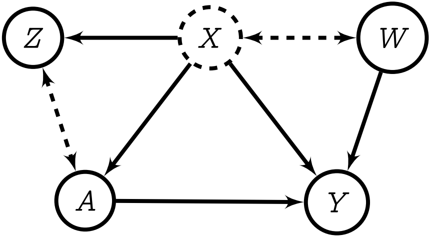
$$E(Y^{(a)}) = \int E(y|x, a)dp(x)$$

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:

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

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

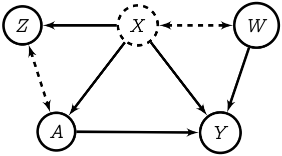
$\Longrightarrow$ Can recover $E(Y^{(a)})$ from observational data!

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

Proof (discrete variables)

If X were observed,

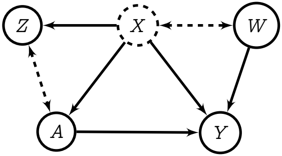
$$P(Y|do(a)) := \sum_{i=1}^{D} P(y|x_i, a)P(x_i)$$



Proof (discrete variables)

If X were observed,

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



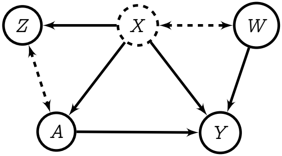
Proof (discrete variables)

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

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



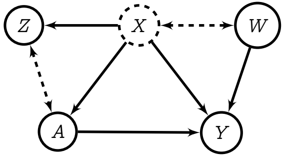
Proof (discrete variables)

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

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



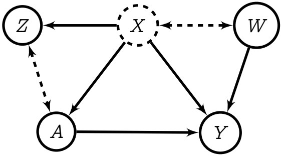
Proof (discrete variables)

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

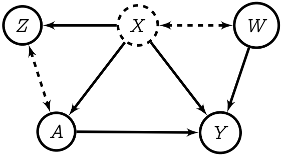
$$P(y|Z, a) = P(y|X, a)P(X|Z, a)$$

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

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

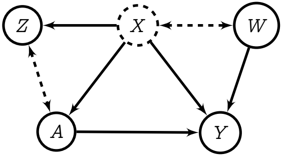
Proof (discrete variables)

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

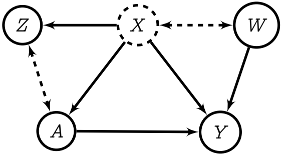
$$P(y|Z, a) = P(y|X, a)\underbrace{P^{-1}(W|X)P(W|Z, a)}_{P(X|Z,a)}$$

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

Proof (discrete variables)

From previous slide:

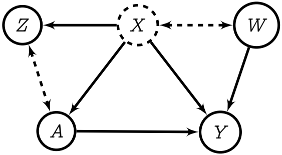
$$p(y|X, a) = p(y|Z, a)P^{-1}(W|Z, a)P(W|X)$$



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 $P(X)$:

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

The proxy correction (continuous)

If X were observed,

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

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

Miao, Geng, Tchetgen Tchetgen (2018)

The proxy correction (continuous)

If X were observed,

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

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

Main theorem: Assume we have solved...

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

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

Miao, Geng, Tchetgen Tchetgen (2018)

The proxy correction (continuous)

If X were observed,

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

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

Main theorem: Assume we have solved...

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

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

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

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

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

Miao, Geng, Tchetgen Tchetgen (2018)

Our solution

- Stage 1: ridge regression from $\phi(a) \otimes \phi(z)$ to $\phi(w)$
  - yields conditional mean embedding $\mu_{W|a,z}$
- Stage 2: ridge regression from $\mu_{W|a,z}$ and $\phi(a)$ to y
  - yields $h_y(w, a)$.
- Solved using sieves [A], kernel [B], or learned NN [C] features

Code available for kernel and NN solutions

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

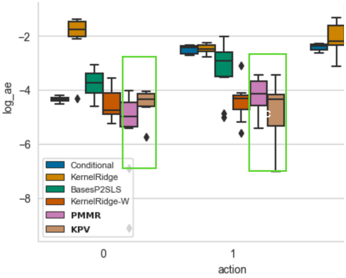
[A] Deaner (2021) Proxy controls and panel data.

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

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

Grade retention and cognitive outcome

- X: unobserved confounder ("ability")
- A: 0: no retention. 1: kindergarten retention. 2: early elementary retention.
- Y: math scores, age 11
- Z: cognitive test scores in elementary school
- W: cognitive test scores from kindergarten



J. Fruehwirth, S. Navarro, Y. Takahashi (2016). How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects.

Deaner (2021)

Conclusions

Kernel ridge regression:

- Solution for ATE, ATT, CATE, mediation analysis, dynamic treatment effects, proximal learning
- ....with treatment A, covariates X, V, mediator M, proxies (W, Z) multivariate, "complicated"
- Simple, robust implementation
- Strong statistical guarantees under general smoothness assumptions

In the papers, but not in this talk:

- Doubly robust estimates for discrete A, V with automatic debiasing
- Elasticities
- Regression to causal effect distributions over Y (not just $E(Y^{(a)}|\ldots)$)
- Instrumental variable regression
- Same algorithms but with adaptive NN features

Selected papers

Unobserved confounders:

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

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

NeurIPS 2019:

arXiv.org > cs > arXiv:1906.00232

Computer Science > Machine Learning

[Submitted on 1 Jun 2019 (v1), last revised 15 Jul 2020 (this version, v6)]

Kernel Instrumental Variable Regression

Rahul Singh, Maneesh Sahani, Arthur Gretton

Observed confounders:

arXiv.org > econ > arXiv:2010.04855

Economics > Econometrics

[Submitted on 10 Oct 2020 (v1), last revised 14 Dec 2021 (this version, v4)]

Generalized Kernel Ridge Regression for Nonparametric Structural Functions and Semiparametric Treatment Effects

Rahul Singh, Liyuan Xu, Arthur Gretton

arXiv.org > stat > arXiv:2111.03950

Statistics > Methodology

[Submitted on 6 Nov 2021]

Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects

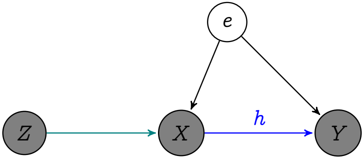
Rahul Singh, Liyuan Xu, Arthur Gretton

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

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

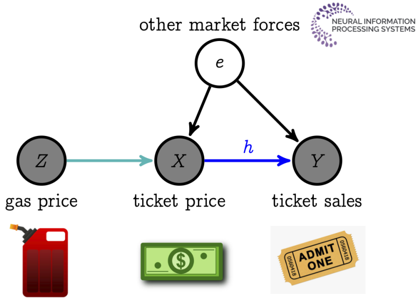


$$Y = \langle h, \psi(X) \rangle + e \qquad \mathbb{E}(e|Z) = 0$$

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

Instrumental variable setting (1)

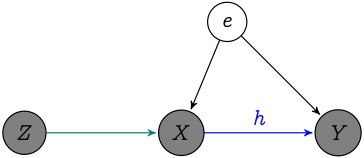
- Unobserved confounder $e \implies$ prediction $\neq$ counterfactual prediction
- goal: learn causal relationship h between input X and output Y
  - if we intervened on X, what would be the effect on Y?
- Instrument Z only influences Y via X, identifying h



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

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)