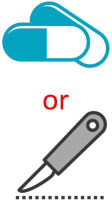
# Causal modelling with kernels: treatment effects, counterfactuals, mediation, and proxies

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Institute of Applied Mathematics, METU, 2025

# A medical treatment scenario

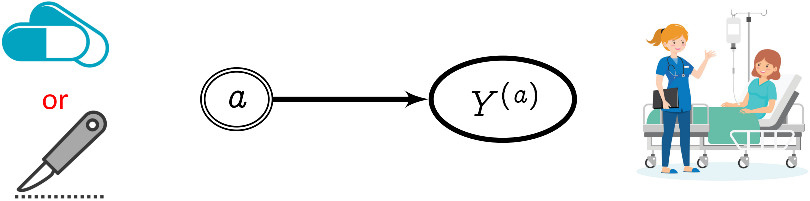


From our observations of historical hospital data:

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

Just recommend pills? Cheaper and more effective!

# A medical treatment scenario



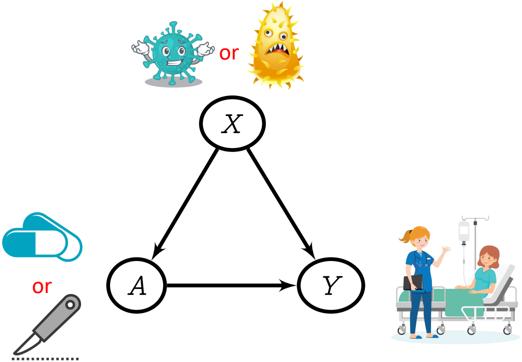
From our intervention (making all patients take a treatment):

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

What went wrong?

# Observation vs intervention

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

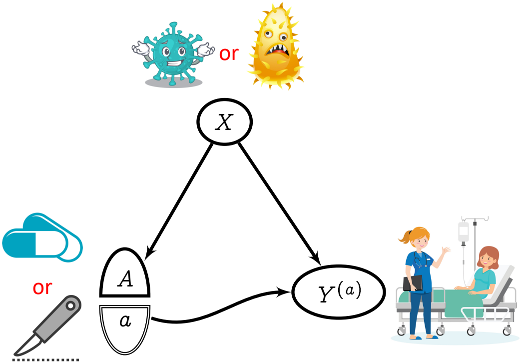


From our <u>observations</u> of historical hospital data:

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

# Observation vs intervention

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



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

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

# Outline

Talk structure:

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

Advantages of the approach:

- Treatment $A$, covariates $X$, etc can be 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 $y_i$:

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

Kernel as feature dot product:

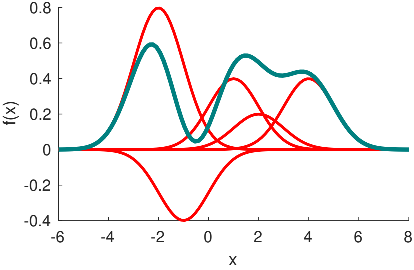
$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$

Solution at $x$:

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

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

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



# Building block: kernel ridge regression

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

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

Kernel as feature dot product:

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

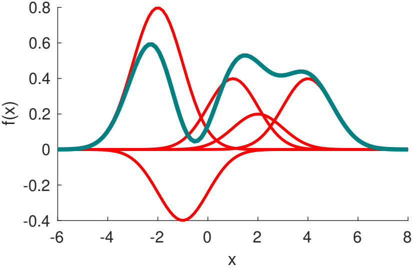
Solution at $x$ (as weighted sum of $y$)

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

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

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

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



# A reminder: the RKHS norm

Eigendecomposition of $k(x,x')$ wrt probability measure $p(x)$,

$$\lambda_\ell e_\ell(x) = \int k(x,x') e_\ell(x') p(x') dx' \qquad \int e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x,x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'),$$

which converges in $L_2(p)$.

Warning: for RKHS, need absolute and uniform convergence, guaranteed via Mercer's theorem under conditions on $p(x)$ and $k(x.x')$.

# A reminder: the RKHS norm

Eigendecomposition of $k(x,x')$ wrt probability measure $p(x)$,

$$\lambda_\ell e_\ell(x) = \int k(x,x') e_\ell(x') p(x') dx' \qquad \int e_i(x) e_j(x) p(x) dx = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

We can write

$$k(x,x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'),$$

which converges in $L_2(p)$.

Warning: for RKHS, need absolute and uniform convergence, guaranteed via Mercer's theorem under conditions on $p(x)$ and $k(x.x')$.

# A reminder: the RKHS norm

For two functions $f, g$ in $L_2(p)$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\langle f, g \rangle_{L_2(p)} = \int_{-\infty}^{\infty} f(x) g(x) p(x) dx = \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell$$

# A reminder: the RKHS norm

For two functions $f, g$ in $L_2(p)$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\langle f, g \rangle_{L_2(p)} = \int_{-\infty}^{\infty} f(x) g(x) p(x) dx = \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell$$

Dot product in $\mathcal{H}$ has roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell}.$$

# A reminder: the RKHS norm

For two functions $f, g$ in $L_2(p)$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{m=1}^{\infty} \hat{g}_m e_m(x),$$

dot product is

$$\langle f, g \rangle_{L_2(p)} = \int_{-\infty}^{\infty} f(x) g(x) p(x) dx = \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell$$

Dot product in $\mathcal{H}$ has roughness penalty,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell}.$$

Define smooth subspace $\mathcal{H}^c \hookrightarrow \mathcal{H} \hookrightarrow L_2(p)$ as

$$\langle f, g \rangle_{\mathcal{H}^c} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell^c} \qquad c > 1.$$

# 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.
- Eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$
  - Larger $b \implies$ easier problem

[A] Fischer, Steinwart (2020). Sobolev norm learning rates for regularized least-squares algorithms.

# 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.
- Eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$
  - Larger $b \implies$ easier problem

Consistency [A, Theorem 1.ii]

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

Best rate is $O_P(n^{-1/4})$ for $c = 2$, $b \to \infty$.

[A] Fischer, Steinwart (2020). Sobolev norm learning rates for regularized least-squares algorithms.

# Observed covariates: (conditional) ATE

Kernels (Biometrika 2023):


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**

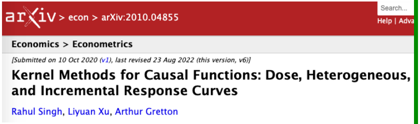
Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:

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

# Observed covariates: (conditional) ATE

## Kernels (Biometrika 2023):



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):

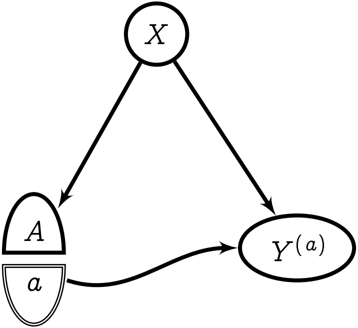
$$\mathrm{E}(Y^{(a)}) = \int E(y|a,x)\,dp(x)$$

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

Assume: (1) 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, ...)



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

# Multiple inputs via products of kernels

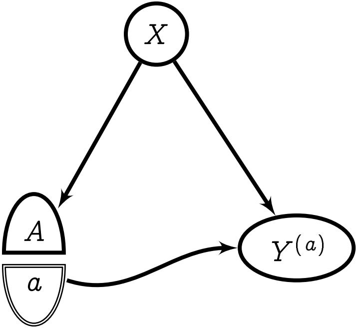
We may predict expected outcome from two inputs

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

Assume we have:

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

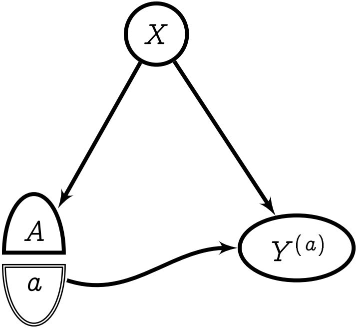
(argument of kernel/feature map indicates feature space)



# 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)

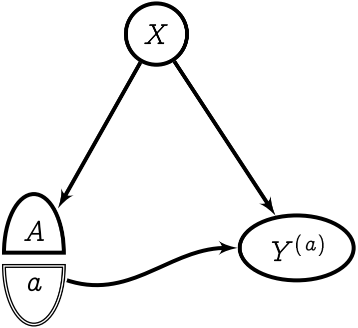
We use outer product of features ( $\implies$ product of kernels):

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

# Multiple inputs via products of kernels

We may predict expected outcome from two inputs

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



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:

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

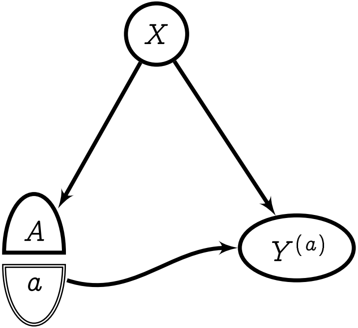
# ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

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



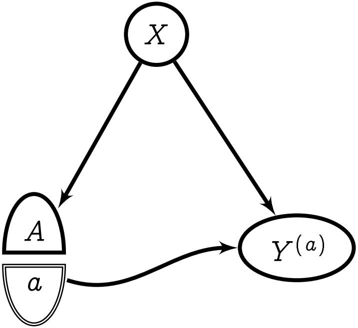
# ATE (dose-response curve)

Well specified setting:

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

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 $P$,

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

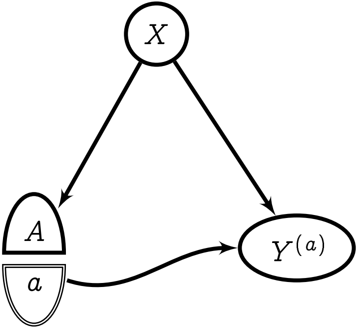
# ATE (dose-response curve)

Well specified setting:

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

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, $\mu_P$ is injective.

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

# ATE: empirical estimate and consistency

Empirical estimate of ATE:

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

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

# 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:

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

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

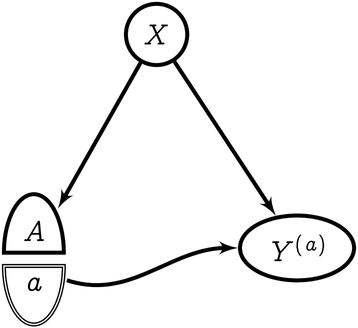
- smoothness assumption $\gamma_0 \in \mathcal{H}^c$, $c \in (1, 2]$
- eigenspectrum decay of input feature covariance, $\eta_j \sim j^{-b}$, $b \geq 1$.

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

# ATE: example

US job corps: training for disadvantaged youths:

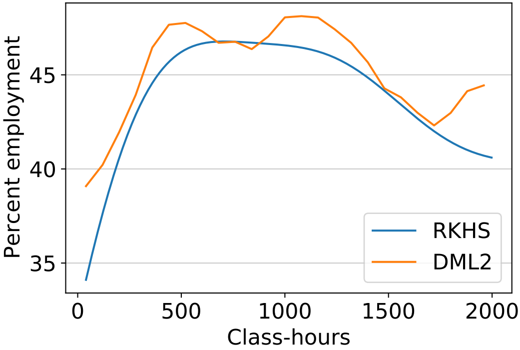
- $X$: covariate/context (age, education, marital status, ...)
- $A$: treatment (training hours)
- $Y$: outcome (percent employment)



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)

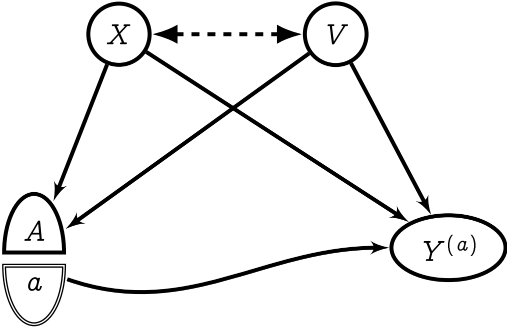
# Conditional average treatment effect

Learned conditional mean:

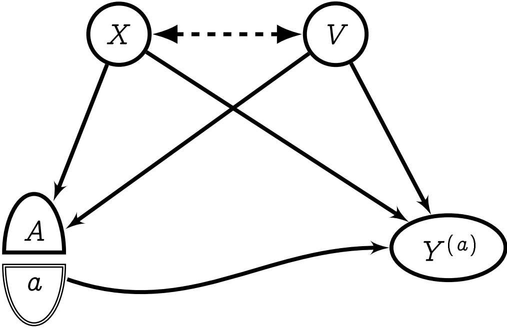
$$\mathrm{E}[\,Y\,|\,a,x,v] \approx \gamma_0(a,x,v)$$
$$= \langle \gamma_0, \varphi(a)\otimes\varphi(x)\otimes\varphi(v)\rangle \,.$$

Conditional ATE

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



# Conditional average treatment effect



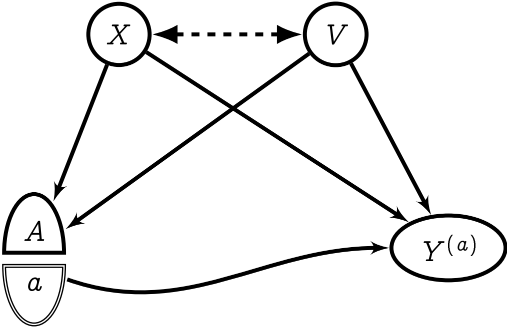
Learned conditional mean:

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

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:

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

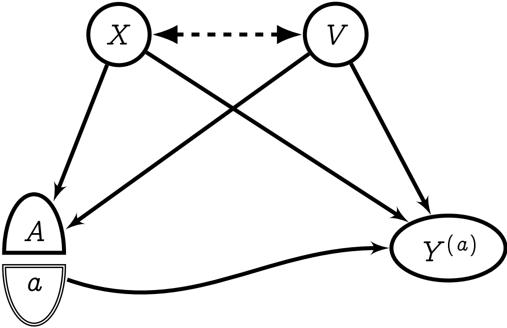
Conditional ATE

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

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

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

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.

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

# Regressing from feature space to feature space

Our goal: an operator $E_0 : \mathcal{H}_\mathcal{V} \rightarrow \mathcal{H}_\mathcal{X}$ such that

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

Assume

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

Implied 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

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

Implied 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

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

Implied 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 $\varphi(x)$:

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

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

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

Assume

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

Implied smoothness assumption:

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

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

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

Ridge regression solution:

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

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

# Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6.3]

$$E_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}^{c_1}, \mathcal{H}_{\mathcal{X}}) \quad c_1 \in (1,2].$$

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

- Eigenspectrum of $\varphi(v)$ covariance decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

[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 6.3]

$$E_0 \in \mathrm{HS}(\mathcal{H}_{\mathcal{V}}^{c_1}, \mathcal{H}_{\mathcal{X}}) \quad c_1 \in (1, 2].$$

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

- Eigenspectrum of $\varphi(v)$ covariance decays as $\eta_{1,j} \sim j^{-b_1}$, $b_1 \geq 1$.

Consistency [A, Theorem 2, Theorem 3]

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

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:

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

Consistency: [A, Theorem 2]

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

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

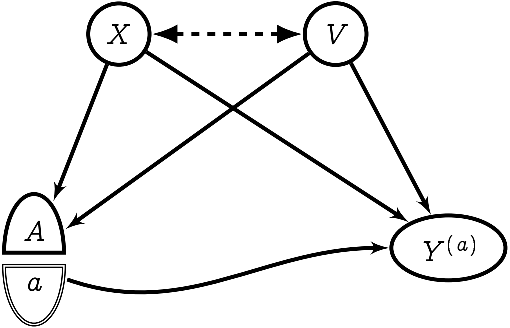
- $E_0 \in \text{HS}(\mathcal{H}_{\mathcal{V}}^{c_1}, \mathcal{H}_{\mathcal{X}})$
- $\gamma_0 \in \mathcal{H}^c$.

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

# Conditional ATE: example

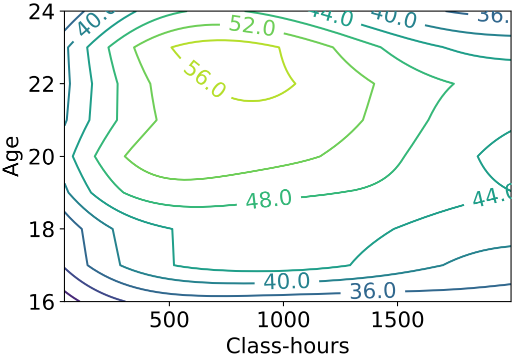
US job corps: training for disadvantaged youths:

- $X$: 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)

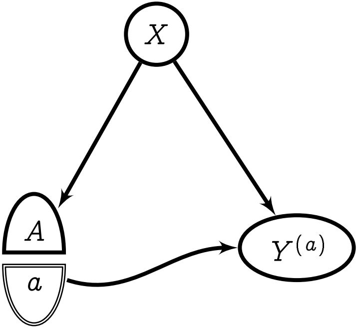
# Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

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



Empirical ATT:

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

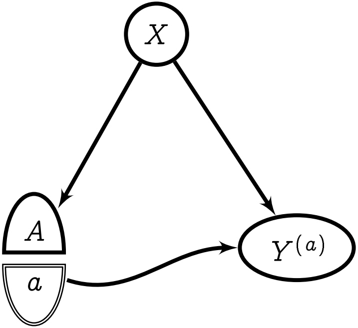
# Counterfactual: average treatment on treated

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:

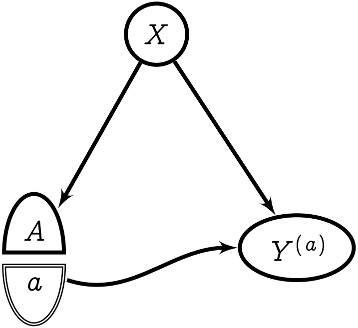
$$\mathrm{E}[Y|a,x] = \gamma_0(a,x)$$

Average treatment on treated:

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

Empirical ATT:

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

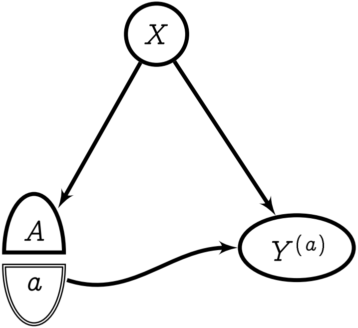


# Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:



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

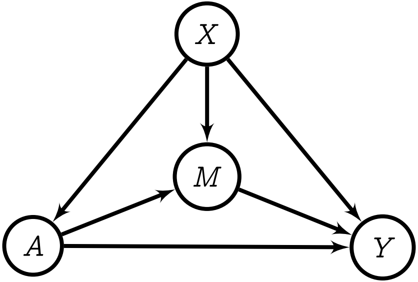
Empirical ATT:

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

# Mediation analysis

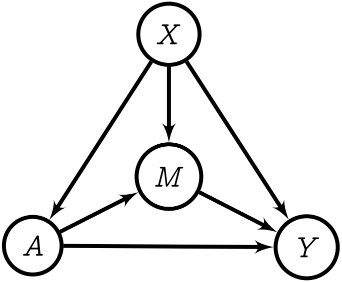
- Direct path from treatment $A$ to effect $Y$
- Indirect path $A \to M \to Y$
- $X$: context

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)

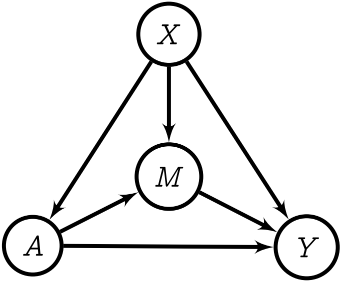
$$\gamma_0(a, m, x) \approx \mathrm{E}[Y | A = a, M = m, X = x]$$

Singh, Xu, G (2022b). 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:

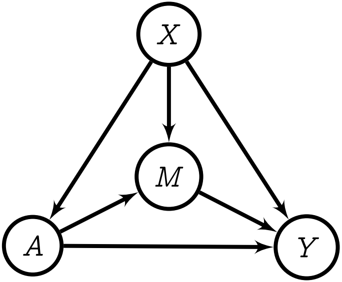
$$Y^{\{a', M^{(a)}\}} = \int \gamma_0(a', M, X) \mathrm{d}\mathbb{P}(M | A = a, X) d\mathbb{P}(X)$$

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

Singh, Xu, G (2022b). 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 $a$

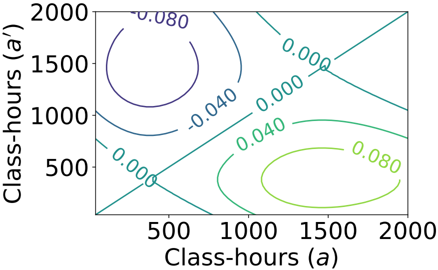
Singh, Xu, G (2022b). Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects.

# Mediation analysis: results

Total effect:

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

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



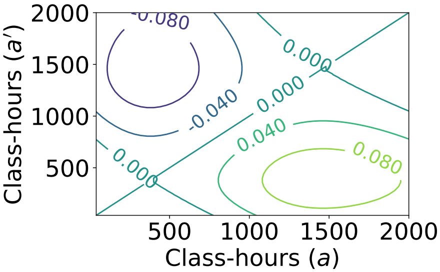
- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests

Singh, Xu, G (2022b)

# Mediation analysis: results

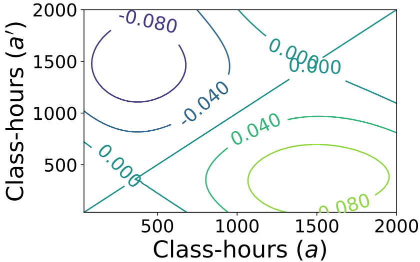
Total effect:

$$\theta_0^{TE}(a, a') := \mathbb{E}[\, Y^{\{a', M^{(a')}\}} - Y^{\{a, M^{(a)}\}}]$$



Direct effect:

$$\theta_0^{DE}(a, a') := \mathbb{E}[\, Y^{\{a', M^{(a)}\}} - Y^{\{a, M^{(a)}\}}]$$

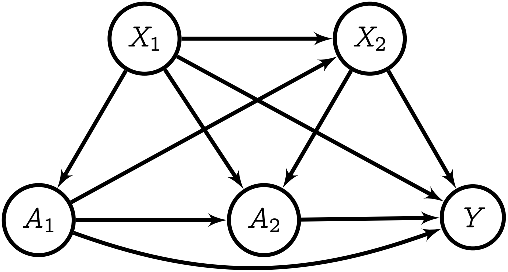


- $a' = 1600$ hours vs $a = 480$ means 0.1 reduction in arrests
- Indirect effect mediated via employment effectively zero

Singh, Xu, G (2022b)

# ...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 $\mathrm{E}(y^{(a_1',a_2')}|A_1 = a_1, A_2 = a_2)$...

(c.f. the Robins G-formula)

# Unobserved confounders: proxy methods

## Kernel features (ICML 2021):

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:

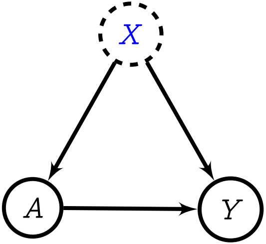
https://github.com/liyuan9988/DeepFeatureProxyVariable/

# 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),

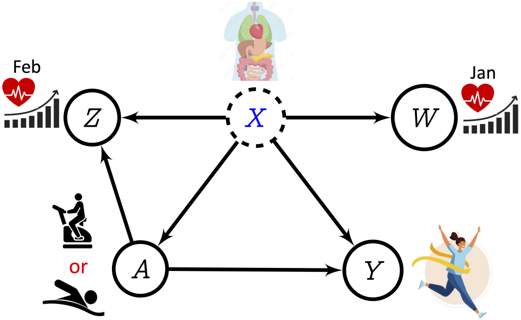
$$\mathrm{E}[Y^{(a)}] = \int \mathrm{E}[Y|X, a]\,dp(X)$$

# 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



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

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

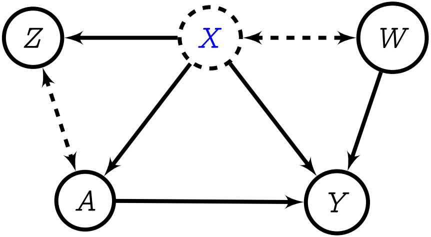
Uehara, Sekhari, Lee, Kallus, Sun (2022) Provably Efficient Reinforcement Learning in Partially Observable Dynamical Systems.

# The proxy correction

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 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.

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 (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 average 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,$^\dagger$ Muandet$^\dagger$ (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

# 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 potential outcome distributions over $Y$ (not just $E(Y^{(a)}|\ldots)$)
- Instrumental variable regression
- Same algorithms but with adaptive NN features

# Selected papers

## Observed confounders:

**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

**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

## Unobserved confounders:

### ICML 2021:

**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:

**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:

**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

# Research support

Work supported by:

The Gatsby Charitable Foundation

GATSBY

Deepmind

Google DeepMind

# Questions?


THE KERNEL
BREWERY LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT