Causal Effect Estimation with Context and Confounders

Arthur Gretton

Gatsby Computational Neuroscience Unit,
Deepmind

Columbia Statistics, 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 \((\text{intervention})\): \(\mathbb{E}[Y^{(a)}] = \sum_x \mathbb{E}[Y|a, x]p(x)\)

From our \textbf{intervention} (making all patients take a treatment):

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

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

Causal effect estimation, observed covariates:

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

Causal effect estimation, hidden covariates:

- ... instrumental variables, proxy variables

What’s new? What is it good for?

- Treatment $A$, covariates $X$, etc can be multivariate, complicated...
- ...by using kernel or adaptive neural net feature representations
Model assumption: linear functions of features

All learned functions will take the form:

\[ \gamma(x) = \gamma^\top \phi(x) = \langle \gamma, \phi(x) \rangle_H \]
Model assumption: linear functions of features

All learned functions will take the form:

\[ \gamma(x) = \gamma^\top \varphi(x) = \langle \gamma, \varphi(x) \rangle_H \]

**Option 1:** Finite dictionaries of learned neural net features \( \varphi_\theta(x) \)
(linear final layer \( \gamma \))

Xu, G., A Neural mean embedding approach for back-door and front-door adjustment. (ICLR 23)
Xu, Chen, Srinivasan, de Freitas, Doucet, G. Learning Deep Features in Instrumental Variable Regression. (ICLR 21)

**Option 2:** Infinite dictionaries of fixed kernel features:

\[ \langle \varphi(x_i), \varphi(x) \rangle_H = k(x_i, x) \]

Kernel is feature dot product.

Singh, Xu, G. Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. (Biometrika 23)
Singh, Sahani, G. Kernel Instrumental Variable Regression. (NeurIPS 19)
Model fitting: ridge regression

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

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

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

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

Neural net solution at $x$:

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

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

$$C_{XX} = \frac{1}{n} \sum_{i=1}^{n} [\varphi(x_i) \varphi(x_i)^\top]$$
Model fitting: ridge regression

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

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

Kernel solution at $x$

(as weighted sum of $y$)

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

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

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

$$(k_{XX})_i = k(x_i, x)$$
**KRR: consistency in RKHS norm**

Assume problem well specified

- Denote: $\gamma_0 \in \mathcal{H}^c$ where $\mathcal{H}^c \subset \mathcal{H}$, $c \in (1, 2]$
  - 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

---

KRR: consistency in RKHS norm

Assume problem well specified

- Denote: $\gamma_0 \in \mathcal{H}^c$ where $\mathcal{H}^c \subset \mathcal{H}$, $c \in (1, 2)$
  - 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 $c = 2$, $b \to \infty$.

Observed covariates: (conditional) ATE

Kernel features (Biometrika 2023):

NN features (ICLR 2023):

A Neural Mean Embedding Approach for Back-door and Front-door Adjustment
Liyuan Xu, Arthur Gretton

Code for NN and kernel causal estimation with observed covariates:
https://github.com/liyuan9988/DeepFrontBackDoor/
Observed covariates: (conditional) ATE

Kernel features (Biometrika 2023):

NN features (ICLR 2023):

Code for NN and kernel causal estimation with observed covariates:
https://github.com/liyuan9988/DeepFrontBackDoor/
Average treatment effect

Potential outcome (intervention):

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

(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 \(Y^{(a)} \perp A \mid X\). (3) Overlap.

**Example:** US job corps, training for disadvantaged youths:

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

\[ X \]
\[ A \]
\[ a \]
\[ Y^{(a)} \]
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):

$$\phi(x, a) = \varphi(a) \otimes \varphi(x) \quad \mathcal{K}([a, x], [a', x']) = k(a, a')k(x, x')$$
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 (\( \Rightarrow \) product of kernels):

\[ \phi(x, a) = \varphi(a) \otimes \varphi(x) \quad \mathcal{K}([a, x], [a', x']) = k(a, a')k(x, x') \]

Ridge regression solution:

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

Well-specified setting:

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

ATE as feature space dot product:

\[ \text{ATE}(a) = \mathbb{E}[\gamma_0(a, X)] \]
\[ = \mathbb{E}[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle] \]
**ATE (dose-response curve)**

Well-specified setting:

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

ATE as feature space dot product:

\[
\text{ATE}(a) = \mathbb{E}[\gamma_0(a, X)]
\]
\[
= \mathbb{E}[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle]
\]
\[
= \langle \gamma_0, \varphi(a) \otimes \mathbb{E}[\varphi(X)] \rangle
\]

Feature map of probability \( P(X) \),

\[
\mu_X = [\ldots \mathbb{E}[\varphi_i(X)] \ldots]
\]
**ATE: example**

US job corps: training for disadvantaged youths:

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

**Empirical ATE:**

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


Singh, Xu, G (2022a).
First 12.5 weeks of classes confer employment gain: from 35% to 47%.

[RKHS] is our $\text{ATE}(a)$.


Singh, Xu, G (2022a)
Conditional average treatment effect

Well-specified setting:

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

Conditional ATE

\[ \text{CATE}(a, v) \]
\[ = \mathbb{E} \left[ Y^{(a)} | V = v \right] \]
Conditional average treatment effect

Well-specified setting:

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

Conditional ATE

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

Well-specified setting:

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

Conditional ATE

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

How to take conditional expectation?

Density estimation for \( p(X|V = v) \)? Sample from \( p(X|V = v) \)?
 Conditional average treatment effect

Well-specified setting:

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

Conditional ATE

CATE($a, v$)

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

Learn conditional mean embedding: $\mu_{X | V = v} := \mathbb{E}_X[\varphi(X) | V = v]$
Regressing from feature space to feature space

Our goal: an operator \( F_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X \) such that

\[
F_0 \varphi(v) = \mu X \big|_{V=v}
\]

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.
Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning.
Regressing from feature space to feature space

Our goal: an operator \( F_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X \) such that

\[
F_0 \varphi(v) = \mu X | V = v
\]

Assume

\[
F_0 \in \text{span} \{ \varphi(x) \otimes \varphi(v) \} \iff F_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)
\]

Implied smoothness assumption:

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

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


Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning
Regressing from feature space to feature space

Our goal: an operator \( F_0 : \mathcal{H}_V \to \mathcal{H}_X \) such that

\[
F_0 \varphi(v) = \mu_X | V = v
\]

Assume

\[
F_0 \in \text{span} \{ \varphi(x) \otimes \varphi(v) \} \iff F_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)
\]

Implied smoothness assumption:

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

A Smooth Operator

Song, Huang, Smola, Fukumizu (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems.
Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning
Regressing from feature space to feature space

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

$$F_0 \varphi(v) = \mu x | v = v$$

Assume

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

Implied smoothness assumption:

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

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

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

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


Li, Meunier, Mollenhauer, G (2022), Optimal Rates for Regularized Conditional Mean Embedding Learning
Regressing from feature space to feature space

Our goal: an operator $F_0 : \mathcal{H}_V \to \mathcal{H}_X$ such that

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

Assume

$$F_0 \in \text{span}\{ \varphi (x) \otimes \varphi (v) \} \iff F_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)$$

Implied smoothness assumption:

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

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

$$\hat{F} = \arg\min_{F} \sum_{\ell=1}^{n} \| \varphi (x_{\ell}) - F \varphi (v_{\ell}) \|^2_{\mathcal{H}_X} + \lambda_2 \| F \|^2_{HS}$$

Ridge regression solution:

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

$$\beta (v) = [K_{VV} + \lambda_2 I]^{-1} k_{V v}$$
Consistency of conditional mean embedding

Assume problem well specified [B, Assumption 6]

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

\( T_1 \) is covariance of features \( \varphi(\nu) \):

- 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).
Consistency of conditional mean embedding

Assume problem well specified \([B, \text{Assumption 6}]\)

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

\(T_1\) is covariance of features \(\varphi(\nu)\):

- 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 \([A, \text{Theorem 2, Theorem 3}]\)

\[
\| \hat{E} - E_0 \|_{HS} = O_P \left( n^{-1 \over 2} \frac{c_1-1}{c_1 + 1/b_1} \right),
\]

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).

Consistency of CATE

Empirical CATE:

\[ \hat{\theta}^{\text{CATE}}(a, v) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot K_{XX} (K_{VV} + n\lambda_1 I)^{-1} K_{Vv} \odot K_{Vv}) \]

from \( \hat{\mu}_{X|V=v} \)
Consistency of CATE

Empirical CATE:

\[ \hat{\theta}^{\text{CATE}}(a, \nu) = Y^\top (K_{AA} \odot K_{XX} \odot K_{VV} + n\lambda I)^{-1} (K_{Aa} \odot K_{XX}(K_{VV} + n\lambda I)^{-1} K_{Vv} \odot K_{Vv}) \]

from \( \hat{\mu}_{X|V=v} \)

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 \( \hat{\mathcal{E}} \) 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 (education, marital status, ...)
- $A$: treatment (training hours)
- $Y$: outcome (percent employed)
- $V$: age

Singh, Xu, G (2022a)
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:

\[ \hat{\theta}_{ATT}(a, a') \]
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

\[ \theta^{\text{ATT}}(a, a') \]

\[ = E[y(a') | A = a] \]

Empirical ATT:

\[ \hat{\theta}^{\text{ATT}}(a, a') \]
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

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

$$= \mathbb{E}[y^{(a')}|A = a]$$

$$= \mathbb{E}_P[\langle \gamma_0, \varphi(a') \otimes \varphi(X) \rangle | A = a]$$

$$= \langle \gamma_0, \varphi(a') \otimes \mathbb{E}_P[\varphi(X)|A = a] \rangle$$

$$\mu_{X|A=a}$$

Empirical ATT:

$$\hat{\theta}^{ATT}(a, a')$$
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

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

Empirical ATT:

\[ \hat{\theta}^{ATT}(a, a') \]
\[ = Y^T (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot K_{XX} (K_{AA} + n\lambda I)^{-1} K_{Aa}) \]

from \( \hat{\mu}_{X | A = a} \)
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 $M$?
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)

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

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)

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

A quantity of interest, the mediated effect:

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

Effect of intervention \( a' \), with \( M(a) \) as if intervention were \( a \)

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)

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

A quantity of interest, the mediated effect:

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

\[ = \langle \gamma_0, \varphi(a') \rangle \otimes \mathbb{E}_P\{\mu_{M|A=a,X} \otimes \varphi(X)\} \]

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

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:

\[ \theta_{TE}^T(a, a') \]

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

Direct effect:

\[ \theta_{DE}^D(a, a') \]

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

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

Singh, Xu, G (2022b)
Dynamic treatment effect: sequence $A_1, A_2$ of treatments.

- potential outcomes $Y^{(a_1)}$, $Y^{(a_2)}$, $Y^{(a_1, a_2)}$,
- counterfactuals $\mathbb{E} \left[ Y^{(a'_1, a'_2)} | A_1 = a_1, A_2 = a_2 \right]$

(c.f. the Robins G-formula)

Conclusions

Neural net and kernel solutions:

- ...for ATE, CATE, dynamic treatment effects
- ...with treatment $A$, covariates $X, V$, proxies $(W, Z)$ multivariate, “complicated”
- Convergence guarantees for kernels and NN

Next lecture:

- Unobserved covariates/confounders (IV and proxy methods)

Code available for all methods
Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind
Questions?