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

Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London

Oxford, 2022
A medical treatment scenario

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?
Observation vs intervention

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

From our observations of historical hospital data:

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

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

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

- $P(Y = \text{cured}|do(\text{pills})) = 0.64$
- $P(Y = \text{cured}|do(\text{surgery})) = 0.75$
Questions we will solve
Outline

Talk structure:

- Average treatment effect (ATE)
  - ...via kernel mean embedding (marginalization)
- Conditional average treatment effect (CATE)
  - via kernel conditional mean embedding
- 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)


**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 $\gamma_0(x) := \mathbb{E}[Y|X = x]$ 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 as feature dot product:

$$\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{H}} = k(x_i, x)$$
Building block: kernel ridge regression

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

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

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

$$(K_{XX})_{ij} = k(x_i, x_j),$$
Building block: kernel ridge regression

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

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

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 \( y \))

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

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

\[
(K_{XX})_{ij} = k(x_i, x_j)
\]

\[
(k_{XX})_i = k(x_i, x)
\]
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, $\eta_j \sim j^{-b}, \quad b \geq 1$
  - 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 $c = 2$, $b \to \infty$.

(Conditional) average treatment effect, average treatment on treated
Average treatment effect

Potential outcome (intervention):

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

(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

\[ \gamma_0(a, x) := E[Y|a, x] \]

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 (\( \otimes \) product of kernels):

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

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

Well specified setting:

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

ATE as feature space dot product:

\[
\theta_0^{ATE}(a) = \mathbb{E}_P[\gamma_0(a, X)] \\
= \mathbb{E}_P \langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle
\]
ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

\[
\theta_0^{ATE}(a) = \mathbb{E}_P[\gamma_0(a, X)] \\
= \mathbb{E}_P \langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle \\
= \langle \gamma_0, \varphi(a) \otimes \mu_P \rangle \\
\quad \underbrace{\mathbb{E}_P \varphi(X)}
\]

Feature map of probability \( P \),

\[ \mu_P = [\ldots \mathbb{E}_P [\varphi_i(X)] \ldots] \]
ATE (dose-response curve)

Well specified setting:

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

ATE as feature space dot product:

\[
\theta_0^{ATE}(a) = \mathbb{E}_P[\gamma_0(a, X)] \\
= \mathbb{E}_P \langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle \\
= \langle \gamma_0, \varphi(a) \otimes \mathbb{E}_P \varphi(X) \rangle \\
= \langle \gamma_0, \varphi(a) \otimes \mu_P \rangle \\
\]

For characteristic kernels, \( \mu_P \) is injective.

Consistency: \( ||\hat{\mu}_P - \mu_P||_\mathcal{H} = \mathcal{O}_P(n^{-1/2}) \)
ATE: empirical estimate and consistency

Empirical estimate of ATE:

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

ATE: empirical estimate and consistency

Empirical estimate of ATE:

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

Consistency:

\[
\left\| \hat{\theta}_{\text{ATE}} - \theta_{\text{ATE}}^0 \right\|_\infty = O_P \left(n^{-\frac{1}{2} \frac{c-1}{c+1/b}} \right)
\]

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, \(\eta_j \sim j^{-b}, \ b \geq 1\).

ATE: example

US job corps: training for disadvantaged youths:

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

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

[RKHS] is our $\hat{\theta}_{ATE}(a)$


Singh, Xu, G (2022a)
Learned conditional mean:

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

Conditional ATE

\[ \theta_{o}^{\text{ATE}}(a, v) = E(Y^{(a)} | V = v) \]
Conditional average treatment effect

Learned conditional mean:

\[
E[Y \mid a, x, v] \approx \gamma_0(a, x, v) \\
= \langle \gamma_0, \varphi(a) \otimes \varphi(x) \otimes \varphi(v) \rangle.
\]

Conditional ATE

\[
\theta_o^{\text{ATE}}(a, v) \\
= E(Y^{(a)} \mid V = v) \\
= E_F(\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle \mid V = v)
\]
Learned conditional mean:

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

Conditional ATE

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

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

Conditional ATE
\[
\theta^\text{CATE}_o(a, v) = E(Y^{(a)}|V = v) = E_P(\langle \gamma_0, \varphi(a) \otimes \varphi(X) \otimes \varphi(V) \rangle | V = v)
\]
\[
= \langle \gamma_0, \varphi(a) \otimes \mathbb{E}_P[\varphi(X)|V = v] \otimes \varphi(v) \rangle
\]

Learn conditional mean embedding: \( \mu_{X|V=v} := E_P(\varphi(X)|V = v) \)
Regressing from feature space to feature space

Our goal: an operator $E_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ 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.


Regressing from feature space to feature space

Our goal: an operator $E_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ such that

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

Assume

$$E_0 \in \text{span} \{ \varphi(x) \otimes \varphi(v) \} \iff E_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)$$

Implied smoothness assumption:

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

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


Regressing from feature space to feature space

Our goal: an operator $E_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ such that

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

Assume

$$E_0 \in \text{span} \{ \varphi(x) \otimes \varphi(v) \} \iff E_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)$$

Implied smoothness assumption:

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

A Smooth Operator

---

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


Regressing from feature space to feature space

Our goal: an operator $\mathcal{E}_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ such that

$$\mathcal{E}_0 \varphi(v) = \mu_X|_{V=v}$$

Assume

$$\mathcal{E}_0 \in \text{span} \{\varphi(x) \otimes \varphi(v)\} \iff \mathcal{E}_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)$$

Implied smoothness assumption:

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

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

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

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


Regressing from feature space to feature space

Our goal: an operator $E_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ such that

$$E_0 \varphi(v) = \mu_X|_{V=v}$$

Assume

$$E_0 \in \text{span}\{\varphi(x) \otimes \varphi(v)\} \iff E_0 \in \text{HS}(\mathcal{H}_V, \mathcal{H}_X)$$

Implied smoothness assumption:

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

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

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

Ridge regression solution:

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

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

Assume problem well specified [A, Assumption 6]

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

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

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

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

---

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

Earlier consistency proofs (with slower rates or finite dimensional $\varphi(x)$):

Singh, Sahani, G (2019)
Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).
Consistency of conditional mean embedding

Assume problem well specified [A, Assumption 6]

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

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

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

Larger \( c_1 \implies \) smoother \( E_0 \implies \) easier problem.

Consistency [A, Proposition 3]

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

best rate is \( O_P(n^{-1/4}) \), and is minimax optimal.

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

Earlier consistency proofs (with slower rates or finite dimensional \( \varphi(x) \)):
Singh, Sahani, G (2019)
Grunewalder, Lever, Baldassarre, Patterson, G, Pontil (2012).
Consistency of CATE

Empirical CATE:

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

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

Empirical CATE:

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

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 \( \hat{E} \) and \( \hat{\gamma} \), under the assumptions:

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

[A] Singh, Xu, G (2022a)
Conditional ATE: example

US job corps: training for disadvantaged youths:

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

Singh, Xu, G (2022a)
Conditional ATE: results

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

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

Singh, Xu, G (2022a)
Counterfactual: average treatment on treated

Conditional mean:

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

Average treatment on treated:

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

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:

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

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:

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

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:

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

Empirical ATT:

\[ \hat{\theta}^{ATT}(a, a') \]
\[ = Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot K_{XX}(K_{AA} + n\lambda_1 I)^{-1} K_{Aa}) \]
\[ \quad \mu_{X|A=a} \]
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:

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

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

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:

\[ \theta_{0}^{TE}(a, a') \]

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

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: sequence $A_1, A_2$ of treatments.

- potential outcomes $Y^{(a_1)}, Y^{(a_2)}, Y^{(a_1,a_2)}$,
- counterfactuals $E(y^{(a'_1,a'_2)}|A_1 = a_1, A_2 = a_2)$...

(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:

\[ W \perp (Z, A) | X \]
\[ Y \perp Z | (A, X) \]

\[ \implies \text{Can recover } E(Y^{(a)}) \text{ from observational data!} \]

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

If $X$ were observed,

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

....but we do not see $p(x)$.
The proxy correction

If $X$ were observed,

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

....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

If $X$ were observed,

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

....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/


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

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

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:

- 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:

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves
Rahul Singh, Liyuan Xu, Arthur Gretton

Kernel Methods for Multistage Causal Inference: Mediation Analysis and Dynamic Treatment Effects
Rahul Singh, Liyuan Xu, Arthur Gretton

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

Unobserved confounders:

ICML 2021:
Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction
Afsaneh Mastouri, Yuchen Zhu, Limor Guitchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, Krikamol Muandet

NeurIPS 2021:
Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation
Liyuan Xu, Heishiro Kanagawa, Arthur Gretton

NeurIPS 2019:
Kernel Instrumental Variable Regression
Rahul Singh, Maneesh Sahani, Arthur Gretton
Questions?
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 $h$

\[
Y = \langle h, \psi(X) \rangle + e \quad \mathbb{E}(e \mid Z) = 0
\]
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 $h$

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

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