Causal Effect Estimation with Context and Confounders

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

AISTATS, 2023
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):

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

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:

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

All learned functions will take the form:

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

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:

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

Kernel is feature dot product.

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

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

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

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

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

$$(k_{XX})_i = k(x_i, x)$$
Observed covariates: (conditional) ATE

Kernel features (in revision, Biometrika):

NN features (ICLR 2023):

[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

Code for NN and kernel causal estimation with observed covariates:

https://github.com/liyuan9988/DeepFrontBackDoor/
Observed covariates: (conditional) ATE

Kernel features (in revision, Biometrika):

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):

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

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

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

- \(A\): treatment (training hours)
- \(Y\): outcome (percentage employment)
- \(X\): covariates (age, education, marital status, …)
Multiple inputs via products of kernels

We may predict expected outcome from two inputs

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

Assume we have:

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

(argument of kernel/feature map indicates feature space)
Multiple inputs via products of kernels

We may predict expected outcome from two inputs

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

Assume we have:

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

(Argument of kernel/feature map indicates feature space)

We use outer product of features (\( \otimes \) product of kernels):

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

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

\[
\begin{align*}
\text{ATE}(a) &= \mathbb{E}[\gamma_0(a, X)] \\
&= \mathbb{E}[\langle \gamma_0, \varphi(a) \otimes \varphi(X) \rangle] \\
&= \langle \gamma_0, \varphi(a) \otimes \mathbb{E}[\varphi(X)] \rangle \\
&= \langle \gamma_0, \varphi(a) \otimes \mu_X \rangle \\
\end{align*}
\]

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:**

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


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:

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

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

Conditional ATE

\[ \text{CATE}(a, v) \]

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

Well-specified setting:

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

Conditional ATE

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

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

Conditional ATE

CATE\((a, v)\)

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

Learn conditional mean embedding: \(\mu_X|V=v := E_X[\varphi(X)|V = v]\)
Regressing from feature space to feature space

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

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

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 $F_0 : \mathcal{H}_\mathcal{V} \rightarrow \mathcal{H}_\mathcal{X}$ 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:

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

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:

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

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 $F_0 : \mathcal{H}_V \to \mathcal{H}_X$ 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 $\varphi(x)$:

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

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 $F_0 : \mathcal{H}_V \rightarrow \mathcal{H}_X$ 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:

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

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

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

Ridge regression solution:

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

$$\beta(v) = [K_{VV} + \lambda_2 I]^{-1} k_{V,v}$$
Conditional ATE: example

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

Empirical CATE:

\[
\widehat{\text{CATE}}(a, v) = \langle \hat{\gamma}_0, \varphi(a) \otimes \hat{F} \varphi(v) \otimes \varphi(v) \rangle \\
\widehat{E}[\varphi(X)|V=v] 
\]

(with consistency guarantees: see paper!)

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)
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)

What if there are hidden confounders?
Illustration: ticket prices for air travel

Ticket price $A$, seats sold $Y$.

What is the effect on seats sold $Y^{(a)}$ of intervening on price $a$?

Illustration: ticket prices for air travel

Ticket price $A$, seats sold $Y$.

What is the effect on seats sold $Y^{(a)}$ of intervening on price $a$?

Illustration: ticket prices for air travel

Unobserved variable $\varepsilon =$desire for travel, affects both price (via airline algorithms) and seats sold.

- Design for travel:
  $\varepsilon \sim N(\mu, 0.1)$
  $\mu \sim U\left\{ -\frac{1}{2}, 0, \frac{1}{2} \right\}$
Illustration: ticket prices for air travel

Unobserved variable $\varepsilon = \text{desire for travel}$, affects both price (via airline algorithms) and seats sold.

- **Desire for travel:**
  
  $\varepsilon \sim \mathcal{N}(\mu, 0.1)$
  
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$

- **Price:**
  
  $A = \varepsilon + Z,$
  
  $Z \sim \mathcal{N}(5, 0.04)$
Illustration: ticket prices for air travel

Unobserved variable $\varepsilon =$ desire for travel, affects both price (via airline algorithms) and seats sold.

- Desire for travel:
  \[ \varepsilon \sim \mathcal{N}(\mu, 0.1) \]
  \[ \mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\} \]

- Price:
  \[ A = \varepsilon + Z, \]
  \[ Z \sim \mathcal{N}(5, 0.04) \]

- Seats sold:
  \[ Y = 10 - A + 2\varepsilon \]
Illustration: ticket prices for air travel

Unobserved variable $\epsilon =$desire for travel, affects both price (via airline algorithms) and seats sold.

- Desire for travel:
  $\epsilon \sim \mathcal{N}(\mu, 0.1)$
  $\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$

- Price:
  $A = \epsilon + Z$, 
  $Z \sim \mathcal{N}(5, 0.04)$

- Seats sold:
  $Y = 10 - A + 2\epsilon$

Average treatment effect:

$$\text{ATE}(a) = \mathbb{E}[Y^{(a)}] = \int (10 - a + 2\epsilon) \, dp(\epsilon) = 10 - a$$
Illustration: ticket prices for air travel

Unobserved variable $\epsilon =$ desire for travel, affects both price (via airline algorithms) and seats sold.

- Desire for travel:
  $$\epsilon \sim \mathcal{N}(\mu, 0.1)$$
  $$\mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}$$

- Price:
  $$A = \epsilon + Z,$$
  $$Z \sim \mathcal{N}(5, 0.04)$$

- Seats sold:
  $$Y = 10 - A + 2\epsilon$$

$Z$ is an instrument (cost of fuel). Condition on $Z$,

$$\mathbb{E}[Y|Z] = 10 - \mathbb{E}[A|Z] + 2\mathbb{E}[\epsilon|Z] = 0$$
Illustration: ticket prices for air travel

Unobserved variable $\varepsilon = \text{desire for travel}$, affects both price (via airline algorithms) and seats sold.

- **Desire for travel:**
  \[ \varepsilon \sim N(\mu, 0.1) \]
  \[ \mu \sim U\left\{ -\frac{1}{2}, 0, \frac{1}{2} \right\} \]

- **Price:**
  \[ A = \varepsilon + Z, \]
  \[ Z \sim N(5, 0.04) \]

- **Seats sold:**
  \[ Y = 10 - A + 2\varepsilon \]

$Z$ is an instrument (cost of fuel). Condition on $Z$,

\[
\mathbb{E}[Y|Z] = 10 - \mathbb{E}[A|Z] + 2\mathbb{E}[\varepsilon|Z] = 0
\]

Regressing from $\mathbb{E}[A|Z]$ to $\mathbb{E}[Y|Z]$ recovers ATE!
The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 was divided, one half awarded to David Card "for his empirical contributions to labour economics", the other half jointly to Joshua D. Angrist and Guido W. Imbens "for their methodological contributions to the analysis of causal relationships"
Instrumental variable regression with NN features

Definitions:
- \( \varepsilon \): unobserved confounder.
- \( A \): treatment
- \( Y \): outcome
- \( Z \): instrument

Assumptions

\[
\mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon|Z] = 0 \\
\mathbb{P}(A|Z) \text{ not constant in } Z \\
Y \perp Z|A \\
Y = \gamma^\top \phi_\theta(A) + \varepsilon
\]
**Instrumental variable regression with NN features**

Definitions:
- \( \varepsilon \): unobserved confounder.
- \( A \): treatment
- \( Y \): outcome
- \( Z \): instrument

Assumptions

\[
\mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon|Z] = 0
\]
\[
P(A|Z) \quad \text{not constant in } Z
\]
\[
Y \perp Z|A
\]
\[
Y = \gamma^\top \phi_\theta(A) + \varepsilon
\]

**Average treatment effect:**

\[
\text{ATE}(a) = \int \mathbb{E}(Y|\varepsilon, a) dp(\varepsilon) = \gamma^\top \phi_\theta(a)
\]
**Instrumental variable regression with NN features**

**Definitions:**
- $\varepsilon$: unobserved confounder.
- $A$: treatment
- $Y$: outcome
- $Z$: instrument

**Assumptions**

\[
\begin{align*}
\mathbb{E}[\varepsilon] &= 0 \quad \mathbb{E}[\varepsilon|Z] = 0 \\
\mathbb{P}(A|Z) & \text{ not constant in } Z \\
Y & \perp Z|A \\
Y &= \gamma^\top \phi_\theta(A) + \varepsilon
\end{align*}
\]

**Average treatment effect:**

\[
\text{ATE}(a) = \int \mathbb{E}(Y|\varepsilon, a) \, d\rho(\varepsilon) = \gamma^\top \phi_\theta(a)
\]

**IV regression:** Condition both sides on $Z$,

\[
\mathbb{E}[Y|Z] = \gamma^\top \mathbb{E}[\phi_\theta(A)|Z] + \mathbb{E}[\varepsilon|Z] = 0
\]
Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):

Kernel Instrumental Variable Regression
Rahul Singh, Maneesh Sahani, Arthur Gretton

Code for NN and kernel IV methods:
https://github.com/liyuan9988/DeepFeatureIV/

NN features (ICLR 2021):

Learning Deep Features in Instrumental Variable Regression
Liyuan Xu, Yutian Chen, Siddharth Srinivasan, Nando de Freitas, Arnaud Doucet, Arthur Gretton
Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):

[Link to NeurIPS 2019 paper]

NN features (ICLR 2021):

[Link to ICLR 2021 paper]

Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/
Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$\mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2$$
IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$
E_{YZ} \left[ (Y - \gamma^T E[\phi_\theta(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2
$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$
E[\phi_\theta(A)|Z] \approx F \phi_\zeta(Z)
$$

with RR loss

$$
E||\phi_\theta(A) - F \phi_\zeta(Z)||^2 + \lambda_1 ||F||_{HS}^2
$$

IV using neural net features

Stage 2 regression (IV): learn NN features \( \phi_\theta(A) \) and linear layer \( \gamma \) to obtain \( Y \) with RR loss:

\[
E_{YZ} \left[ \left( Y - \gamma^T E[\phi_\theta(A)|Z] \right)^2 \right] + \lambda_2||\gamma||^2
\]

Stage 1 regression: learn NN features \( \phi_\zeta(Z) \) and linear layer \( F \):

\[
E[\phi_\theta(A)|Z] \approx F \phi_\zeta(Z)
\]

with RR loss

\[
E||\phi_\theta(A) - F \phi_\zeta(Z)||^2 + \lambda_1||F||_{HS}^2
\]

Challenge: how to learn \( \theta \)?

---

**IV using neural net features**

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$
\mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2
$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$
\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)
$$

with RR loss

$$
\mathbb{E}[||\phi_\theta(A) - F\phi_\zeta(Z)||^2 + \lambda_1 ||F||^2_{HS}]
$$

**Challenge:** how to learn $\theta$?

From Stage 2 regression?
**IV using neural net features**

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$\mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E}||\phi_\theta(A) - F\phi_\zeta(Z)||^2 + \lambda_1 ||F||^2_{HS}$$

**Challenge:** how to learn $\theta$?

From Stage 2 regression?

...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from Stage 1 regression
IV using neural net features

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$
\mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2 \right] + \lambda_2 \|\gamma\|^2
$$

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$
\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)
$$

with RR loss

$$
\mathbb{E}\|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2
$$

Challenge: how to learn $\theta$?

From Stage 2 regression?

...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from Stage 1 regression

...which requires $\phi_\theta(A)$... which requires $\theta$...

**IV using neural net features**

**Stage 2 regression (IV):** learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$
\mathbb{E}_{YZ} \left[ \left( Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z] \right)^2 \right] + \lambda_2 \|\gamma\|^2
$$

**Stage 1 regression:** learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$
\mathbb{E}[\phi_\theta(A)|Z] \approx F \phi_\zeta(Z)
$$

with RR loss

$$
\mathbb{E}\|\phi_\theta(A) - F \phi_\zeta(Z)\|^2 + \lambda_1 \|F\|_{HS}^2
$$

**Challenge:** how to learn $\theta$?

**From Stage 2 regression?**

...which requires $\mathbb{E}[\phi_\theta(A)|Z]$ from **Stage 1 regression**

...which requires $\phi_\theta(A)$... which requires $\theta$...

**Use the linear final layers!** (i.e. $\gamma$ and $F$)

IV using neural net features

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$\mathbb{E}[\phi_\theta(A) | Z] \approx F \phi_\zeta(Z)$$

with RR loss

$$\mathbb{E} \left[ ||\phi_\theta(A) - F \phi_\zeta(Z)||^2 \right] + \lambda_1 ||F||^2_{HS}$$
IV using neural net features

Stage 1 regression: learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E} \left[ \|\phi_\theta(A) - F\phi_\zeta(Z)\|^2 \right] + \lambda_1\|F\|_H^2$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt $\phi_\theta, \phi_\zeta$:

$$\hat{F}_{\theta,\zeta} = C_{AZ}(C_{ZZ} + \lambda_1 I)^{-1}$$

$$C_{AZ} = \mathbb{E}[\phi_\theta(A)\phi_\zeta^\top(Z)]$$

$$C_{ZZ} = \mathbb{E}[\phi_\zeta(Z)\phi_\zeta^\top(Z)]$$
**IV using neural net features**

**Stage 1 regression:** learn NN features $\phi_\zeta(Z)$ and linear layer $F$:

$$\mathbb{E}[\phi_\theta(A)|Z] \approx F\phi_\zeta(Z)$$

with RR loss

$$\mathbb{E} \left[ ||\phi_\theta(A) - F\phi_\zeta(Z)||^2 \right] + \lambda_1 ||F||^2_{HS}$$

$\hat{F}_{\theta,\zeta}$ in closed form wrt $\phi_\theta, \phi_\zeta$:

$$\hat{F}_{\theta,\zeta} = C_{AZ}(C_{ZZ} + \lambda_1 I)^{-1} \quad C_{AZ} = \mathbb{E}[\phi_\theta(A)\phi_\zeta^\top(Z)]$$

$$C_{ZZ} = \mathbb{E}[\phi_\zeta(Z)\phi_\zeta^\top(Z)]$$

Plug $\hat{F}_{\theta,\zeta}$ into S1 loss, take gradient steps for $\zeta$ (...but not $\theta$...)

---

Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$L_2(\gamma, \theta) = \mathbb{E}_{YZ}[(Y - \gamma^T \mathbb{E}[\phi_\theta(A)|Z])^2] + \lambda_2 ||\gamma||^2$$
Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_{\theta}(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$L_2(\gamma, \theta) = \mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_{\theta}(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2$$

$$= \mathbb{E}_{YZ}[ (Y - \gamma^\top \hat{F}_{\theta,\zeta} \phi_{\zeta}(Z))^2 ] + \lambda_2 ||\gamma||^2$$

From linear final layers in Stages 1, 2:
Learn $\phi_{\theta}(A)$ by plugging $\gamma$ into $S_2$ loss, taking gradient steps for $\gamma$...but changes with $\lambda_2$...so alternate first and second stages until convergence.
Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$L_2(\gamma, \theta) = \mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A) | Z])^2 \right] + \lambda_2 \|\gamma\|^2$$

$$= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta} \phi_\zeta(Z))^2] + \lambda_2 \|\gamma\|^2$$

$\hat{\gamma}_\theta$ in closed form wrt $\phi_\theta$:

$$\hat{\gamma}_\theta := \tilde{C}_{YZ}(\tilde{C}_{ZZ} + \lambda_2 I)^{-1}$$

$$\tilde{C}_{YZ} = \mathbb{E} \left[ Y \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right]^\top \right]$$

$$\tilde{C}_{ZZ} = \mathbb{E} \left[ \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right] \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right]^\top \right]$$
Stage 2: IV regression

Stage 2 regression (IV): learn NN features $\phi_\theta(A)$ and linear layer $\gamma$ to obtain $Y$ with RR loss:

$$
\mathcal{L}_2(\gamma, \theta) = \mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_\theta(A)|Z])^2 \right] + \lambda_2 \|\gamma\|^2 \\
= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta} \phi_\zeta(Z))^2] + \lambda_2 \|\gamma\|^2
$$

$\hat{\gamma}_\theta$ in closed form wrt $\phi_\theta$:

$$
\hat{\gamma}_\theta := \overline{C}_{YZ}(\overline{C}_{ZZ} + \lambda_2 I)^{-1} \\
\overline{C}_{YZ} = \mathbb{E} \left[ Y \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right]^\top \right] \\
\overline{C}_{ZZ} = \mathbb{E} \left[ \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right] \left[ \hat{F}_{\theta,\zeta} \phi_\zeta(Z) \right]^\top \right]
$$

From linear final layers in Stages 1,2:
Learn $\phi_\theta(A)$ by plugging $\hat{\gamma}_\theta$ into S2 loss, taking gradient steps for $\theta$
Stage 2: IV regression

Stage 2 regression (IV): learn NN features \(\phi_{\theta}(A)\) and linear layer \(\gamma\) to obtain \(Y\) with RR loss:

\[
\mathcal{L}_2(\gamma, \theta) = \mathbb{E}_{YZ} \left[ (Y - \gamma^\top \mathbb{E}[\phi_{\theta}(A)|Z])^2 \right] + \lambda_2 ||\gamma||^2 \\
= \mathbb{E}_{YZ}[(Y - \gamma^\top \hat{F}_{\theta,\zeta} \phi_{\zeta}(Z))^2] + \lambda_2 ||\gamma||^2
\]

\(\hat{\gamma}_{\theta}\) in closed form wrt \(\phi_{\theta}\):

\[
\hat{\gamma}_{\theta} := \widetilde{C}_{YZ}(\widetilde{C}_{ZZ} + \lambda_2 I)^{-1} \\
\widetilde{C}_{YZ} = \mathbb{E} \left[ Y \left[ \hat{F}_{\theta,\zeta} \phi_{\zeta}(Z) \right]^\top \right] \\
\widetilde{C}_{ZZ} = \mathbb{E} \left[ \left[ \hat{F}_{\theta,\zeta} \phi_{\zeta}(Z) \right] \left[ \hat{F}_{\theta,\zeta} \phi_{\zeta}(Z) \right]^\top \right]
\]

From linear final layers in Stages 1,2:

Learn \(\phi_{\theta}(A)\) by plugging \(\hat{\gamma}_{\theta}\) into S2 loss, taking gradient steps for \(\theta\)

...but \(\zeta\) changes with \(\theta\)

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

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)
Neural IV in reinforcement learning

Policy evaluation: want Q-value:

\[ Q^\pi(s, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, A_0 = a \right] \]

for policy \( \pi(A|S = s) \).

Application of IV: reinforcement learning

Q value is a minimizer of Bellman loss

\[ \mathcal{L}_{\text{Bellman}} = \mathbb{E}_{SAR} \left[ (R + \gamma \mathbb{E} [Q^\pi(S', A') | S, A] - Q^\pi(S, A))^2 \right]. \]

Corresponds to “IV-like” problem

\[ \mathcal{L}_{\text{Bellman}} = \mathbb{E}_{YZ} \left[ (Y - \mathbb{E}[f(X) | Z])^2 \right] \]

with

\[
\begin{align*}
Y &= R, \\
X &= (S', A', S, A) \\
Z &= (S, A), \\
f_0(X) &= Q^\pi(s, a) - \gamma Q^\pi(s', a')
\end{align*}
\]

RL experiments and data:
https://github.com/liyuan9988/IVOPEwithACME


Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)

Results on mountain car problem

Good performance compared with FQE.

Warning: IV assumption can fail when regression underfits. See papers for details.

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)
...but seriously, what if there are hidden confounders?
The proxy correction

Unobserved $\varepsilon$ with (possibly) complex nonlinear effects on $A$, $Y$

The definitions are:

- $\varepsilon$: unobserved confounder.
- $A$: treatment
- $Y$: outcome

If $\varepsilon$ were observed (which it isn’t),

$$
\mathbb{E}[Y^{(a)}] = \int \mathbb{E}[Y|\varepsilon, a] d\rho(\varepsilon)
$$
The proxy correction

Unobserved $\epsilon$ with (possibly) complex nonlinear effects on $A$, $Y$

The definitions are:

- $\epsilon$: 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.

The proxy correction

Unobserved $\varepsilon$ with (possibly) complex nonlinear effects on $A, Y$

The definitions are:
- $\varepsilon$: unobserved confounder.
- $A$: treatment
- $Y$: outcome
- $Z$: treatment proxy
- $W$: outcome proxy

Structural assumption:

$$W \perp (Z, A)|\varepsilon$$
$$Y \perp Z|(A, \varepsilon)$$

$\implies$ 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.

Unobserved confounders: proxy methods

Kernel features (ICML 2021):

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, Kirakom Muandet

NN features (NeurIPS 2021):

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/
Conclusions

Neural net and kernel solutions:

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

Not in this talk:

- Elasticities
- Regression to potential outcome distributions over $Y$ (not just $E(Y^{(a)}|\ldots)$)

Code available for all methods
Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind
Questions?
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:

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

Average treatment on treated:

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

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

Empirical ATT:

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

Conditional mean:

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

Average treatment on treated:

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

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:

\[ \theta_{ATT}(a, a') \]

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

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

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

\[ \mu_{X|A=a} \]

Empirical ATT:

\[ \hat{\theta}_{ATT}(a, a') \]

\[ = Y^\top (K_{AA} \odot K_{XX} + n\lambda I)^{-1} (K_{Aa'} \odot \left( K_{XX}(K_{AA} + n\lambda_1 I)^{-1} K_{Aa} \right) \]

\[ \text{from } \hat{\mu}_{X|A=a} \]