Causal Effect Estimation with Hidden Confounders using Instruments and Proxies

Arthur Gretton

Gatsby Computational Neuroscience Unit
Google Deepmind

ELLIS Robust Machine Learning Workshop, 2023
Introduction: observation vs intervention

Conditioning from observation: \( \mathbb{E}[Y|A = a] = \sum_x \mathbb{E}[Y|a, \epsilon]p(\epsilon|a) \)

From our observations of historical hospital data:

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

Average causal effect (intervention): \( \mathbb{E}[Y^{(a)}] = \sum_\varepsilon \mathbb{E}[Y|a, \varepsilon]p(\varepsilon) \)

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
Questions we will solve
Outline

Causal effect estimation, robust to 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:

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

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

Neural net solution at $x$:

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

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

$$C_{XX} = \frac{1}{n} \sum_{i=1}^{n} [\varphi(x_i) \varphi(x_i)^\top]$$
Instrumental variable regression
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 $\varepsilon = \text{desire for travel}$, affects both price (via airline algorithms) and seats sold.

- **Desire for travel:**
  \[
  \varepsilon \sim \mathcal{N}(\mu, 0.1) \\
  \mu \sim \mathcal{U}\left\{-\frac{1}{2}, 0, \frac{1}{2}\right\}
  \]
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:**
  $A = \varepsilon + Z$
  $Z \sim \mathcal{N}(5, 0.04)$
Illustration: ticket prices for air travel

Unobserved variable $\epsilon = \text{desire for travel}$, 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:**
  \[ Y = 10 - A + 2\epsilon \]
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:**

$A = \varepsilon + Z,$

$Z \sim \mathcal{N}(5, 0.04)$

**Seats sold:**

$Y = 10 - A + 2\varepsilon$

Average treatment effect:

$$\text{ATE}(a) = \mathbb{E}[Y^{(a)}] = \int (10 - a + 2\varepsilon) \, dp(\varepsilon) = 10 - a$$
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:**
  \[ A = \varepsilon + Z, \]
  \[ Z \sim \mathcal{N}(5, 0.04) \]

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

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

\[ \mathbb{E}[Y|Z] = 10 - \mathbb{E}[A|Z] + 2\mathbb{E}[\varepsilon|Z] = 0 \]
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:**
  - $A = \varepsilon + Z$
  - $Z \sim \mathcal{N}(5, 0.04)$
- **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

\[ \mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon|Z] = 0 \]
\( Z \notin A \)
\( (Y \perp Z|A)_{G_A} \)
\( Y = \gamma^\top \phi_\theta(A) + \varepsilon \)
Instrumental variable regression with NN features

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

Assumptions

\[
\mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon|Z] = 0 \\
Z \perp A \\
(Y \perp Z|A)_{G_A} \\
Y = \gamma^\top \phi_\theta(A) + \varepsilon
\]

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

\[
\mathbb{E}[\varepsilon] = 0 \quad \mathbb{E}[\varepsilon|Z] = 0 \\
Z \perp A \\
(Y \perp Z|A)_{G_A} \\
Y = \gamma^\top \phi_\theta(A) + \varepsilon
\]

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 \( Z \),

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

Kernel features (NeurIPS 2019):

NN features (ICLR 2021):

Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/
Two-stage least squares for IV regression

Kernel features (NeurIPS 2019):

[Image 75x355 to 395x432]

[Image 133x242 to 215x341]

[Image 241x242 to 335x336]

[Image 447x347 to 767x430]

[Image 447x233 to 528x329]

[Image 556x233 to 649x326]

[Image 676x233 to 758x333]

[Image 447x128 to 540x221]

[Image 566x128 to 648x231]

[Image 68x536]Two-stage least squares for IV regression

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

[Image 133x242 to 215x341]

[Image 241x242 to 335x336]

[Image 447x347 to 767x430]

[Image 447x233 to 528x329]

[Image 556x233 to 649x326]

[Image 676x233 to 758x333]

[Image 447x128 to 540x221]

[Image 566x128 to 648x231]

[Image 68x536]

[Image 75x452]

NN features (ICLR 2021):

[Image 133x242 to 215x341]

[Image 241x242 to 335x336]

[Image 447x347 to 767x430]

[Image 447x233 to 528x329]

[Image 556x233 to 649x326]

[Image 676x233 to 758x333]

[Image 447x128 to 540x221]

[Image 566x128 to 648x231]

[Image 68x536]

[Image 75x452]

[Image 133x242 to 215x341]

[Image 241x242 to 335x336]

[Image 447x347 to 767x430]

[Image 447x233 to 528x329]

[Image 556x233 to 649x326]

[Image 676x233 to 758x333]

[Image 447x128 to 540x221]

[Image 566x128 to 648x231]

[Image 68x536]

[Image 75x452]

Code for NN and kernel IV methods:

https://github.com/liyuan9988/DeepFeatureIV/
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:

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

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

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

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

---

**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

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

**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 $\theta$...

---

*Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021) Learning Deep Features in Instrumental Variable 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

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

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

Neural IV in reinforcement learning

Policy evaluation: want Q-value:

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

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

$$Y = R,$$

$$X = (S', A', S, A)$$

$$Z = (S, A),$$

$$f_0(X) = Q^\pi(s, a) - \gamma Q^\pi(s', a').$$

**RL experiments and data:**

[https://github.com/liyuan9988/IVOPEwithACME](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)
Proxy/Negative Control Methods
The proxy correction

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

The definitions are:

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

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

$$\mathbb{E}[Y^{(a)}] = \int \mathbb{E}[Y|\epsilon, a] d\rho(\epsilon)$$
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

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, Krikamol Muandet

Code for NN and kernel proxy methods:

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

NN features (NeurIPS 2021):

Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation

Liyuan Xu, Heishiro Kanagawa, Arthur Gretton
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!
Main theorem

If $\varepsilon$ were observed, we would write *(average treatment effect)*

\[ p(y|do(a)) = \int p(y|a, \varepsilon)p(\varepsilon)d\varepsilon. \]

....but we do not observe $\varepsilon$. 
Main theorem

If \( \varepsilon \) were observed, we would write *(average treatment effect)*

\[
p(y|do(a)) = \int p(y|a, \varepsilon)p(\varepsilon)d\varepsilon.
\]

....but we do not observe \( \varepsilon \).

Main theorem: Assume we solved:

\[
p(y|a, z) = \int h_y(w, a)p(w|a, z)dw
\]

Both \( p(y|a, z) \) and \( p(w|a, z) \) are in terms of observed quantities.
Main theorem

If $\varepsilon$ were observed, we would write (average treatment effect)

$$p(y|do(a)) = \int_u p(y|a, \varepsilon)p(\varepsilon)d\varepsilon.$$ 

....but we do not observe $\varepsilon$.

Main theorem: Assume we solved:

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

Both $p(y|a, z)$ and $p(w|a, z)$ are in terms of observed quantities.

Average treatment effect via $p(w)$:

$$p(y|do(a)) = \int h_y(a, w)p(w)dw$$
Proof (1)

Because $W \perp (Z, A) \mid \epsilon$, we have

$$p(w \mid a, z) = \int p(w \mid \epsilon)p(\epsilon \mid a, z) d\epsilon$$
Proof (1)

Because $W \perp (Z, A) | \varepsilon$, we have

$$p(w|a, z) = \int p(w|\varepsilon)p(\varepsilon|a, z)d\varepsilon$$

Because $Y \perp Z|(A, \varepsilon)$ we have

$$p(y|a, z) = \int p(y|a, \varepsilon)p(\varepsilon|a, z)d\varepsilon$$
**Proof (3)**

Given the solution $h_y$ to:

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

(well defined under identifiability conditions for Fredholm equation of first kind)
Proof (3)

Given the solution \( h_y \) to:

\[
p(y|a, z) = \int h_y(w, a)p(w|a, z)\,dw
\]

(well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

\[
\int p(y|a, \epsilon)p(\epsilon|a, z)\,d\epsilon = \int h_y(w, a) \int p(w|\epsilon)p(\epsilon|a, z)\,d\epsilon\,dw
\]
**Proof (3)**

Given the solution \( h_y \) to:

\[
p(y|a, z) = \int h_y(w, a)p(w|a, z)d\omega
\]

(Well defined under identifiability conditions for Fredholm equation of first kind)

From last slide

\[
\int p(y|a, \varepsilon)p(\varepsilon|a, z)d\varepsilon = \int h_y(w, a)\int p(w|\varepsilon)p(\varepsilon|a, z)d\varepsilon d\omega
\]

This implies:

\[
p(y|a, \varepsilon) = \int h_y(w, a)p(w|\varepsilon)d\omega
\]

under identifiability condition

\[
\mathbb{E}[f(\varepsilon)|A = a, Z = z] = 0, \forall (z, a) \iff f(\varepsilon) = 0, \mathbb{P}_{\varepsilon|A=a} \ a.s. \quad (\triangle)
\]
Proof (4)

From last slide,

\[ p(y|a, \varepsilon) = \int h_y(w, a) p(w|\varepsilon) dw \]

Thus

\[ p(y|do(a)) = \int_u p(y|a, \varepsilon) p(\varepsilon) du \]
Proof (4)

From last slide,

\[ p(y|a, \varepsilon) = \int h_y(w, a)p(w|\varepsilon) \, dw \]

Thus

\[ p(y|do(a)) = \int_u p(y|a, \varepsilon)p(\varepsilon) \, du \]

\[ = \int_u \left[ \int h_y(w, a)p(w|\varepsilon) \, dw \right] p(\varepsilon) \, d\varepsilon \]
Proof (4)

From last slide,

\[ p(y|a, \varepsilon) = \int h_y(w, a)p(w|\varepsilon)dw \]

Thus

\[ p(y|do(a)) = \int_u p(y|a, \varepsilon)p(\varepsilon)du \]

\[ = \int_u \left( \int h_y(w, a)p(w|\varepsilon)d\omega \right) p(\varepsilon)d\varepsilon \]

\[ = \int h_y(w, a)p(w)d\omega \]
Feature implementation

Stage 2: minimize

\[ h_{\lambda_2} = \arg \min_{h \in \mathcal{H}} \mathbb{E}_{y, a, z} \left( y - h^\top \left( \mu_{W|a, z} \otimes \phi(a) \right) \right)^2 + \lambda_2 \| h \|^2_{\mathcal{H}} \]

which is conditional feature mean implementation of

\[ p(y|a, z) = \int h_y(w, a) p(w|a, z) dw \]
Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg \min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left( y - h^\top \left( \mu_{W|a,z} \otimes \phi(a) \right) \right)^2 + \lambda_2 \| h \|_{\mathcal{H}}^2$$

which is conditional feature mean implementation of

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

Stage 1: ridge regression

$$F_{\lambda_1} = \arg \min_{F \in HS} \mathbb{E}_{w,a,z} \left\| \phi(w) - F^\top [\phi(a) \otimes \phi(z)] \right\|_{\mathcal{H}_W}^2 + \lambda_1 \| F \|_{HS}^2$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}^\top [\phi(a) \otimes \phi(z)]$$

Deaner (2021).
Feature implementation

Stage 2: minimize

$$h_{\lambda_2} = \arg \min_{h \in \mathcal{H}} \mathbb{E}_{y,a,z} \left( y - h^\top \left( \mu_{W|a,z} \otimes \phi(a) \right) \right)^2 + \lambda_2 \| h \|^2_\mathcal{H}$$

which is conditional feature mean implementation of

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

Stage 1: ridge regression

$$F_{\lambda_1} = \arg \min_{F \in HS} \mathbb{E}_{w,a,z} \left\| \phi(w) - F^\top [\phi(a) \otimes \phi(z)] \right\|^2_{\mathcal{H}_W} + \lambda_1 \| F \|^2_{HS}$$

which gives us

$$\mu_{W|a,z} = F_{\lambda_1}^\top [\phi(a) \otimes \phi(z)]$$

Average treatment effect estimate:

$$\mathbb{E}_y(y|do(a)) = h_{\lambda_2}^\top (\phi(a) \otimes \mu_W),$$

where $\mu_W = \mathbb{E}_W \phi(W)$

Deaner (2021).


Synthetic experiment, adaptive neural net features

dSprite example:

- $\epsilon = \{\text{scale, rotation, posX, posY}\}$
- Treatment $A$ is the image generated (with Gaussian noise)
- Outcome $Y$ is quadratic function of $A$ with multiplicative confounding by $\text{posY}$.
- $Z = \{\text{scale, rotation, posX}\}$, $W = \text{noisy image sharing posY}$
- Comparison with CEVAE (Louzios et al. 2017)

Confounded offline policy evaluation

Synthetic dataset, demand prediction for flight purchase.

- Treatment $A$ is ticket price.
- Policy $A \sim \pi(Z)$ depends on fuel price.
Conclusions

**Neural net and kernel solutions:**

- ...for instrumental variable regression
- ...for proxy methods
- ...with treatment $A$, covariates $X, V$, proxies $(W, Z)$ multivariate, “complicated”
- Convergence guarantees for kernels and NN

**Code available for all methods**
Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind
Questions?
IV regression using neural net features
**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:

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

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

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

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

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

**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:

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

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:

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

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

...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 $F$:

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

with RR loss

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

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

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

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

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

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

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

Stage 1

---

Xu, Chen, Srinivasan, De Freitas, Doucet, G. (2021)
Stage 2: IV regression

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

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

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

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

$$\hat{\gamma}_\theta := \overline{C}_{YA|Z}(\overline{C}_{AA|Z} + \lambda_2 I)^{-1}$$

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

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

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

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

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

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

$$
\hat{\gamma}_\theta := \widetilde{C}_{YA|Z}^{-1} \left( \widetilde{C}_{AA|Z} + \lambda_2 I \right)
$$

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

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

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:

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

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

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

$$\hat{\gamma}_\theta := \mathbb{C}_{YA|Z}(\mathbb{C}_{AA|Z} + \lambda_2 I)^{-1}$$

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

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

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)