Counterfactual Policy Evaluation in Reproducing Kernel Hilbert Spaces

Krikamol Muandet

Max Planck Institute for Intelligent Systems
Tübingen, Germany

Jeju, Korea — February 22, 2019

Acknowledgment

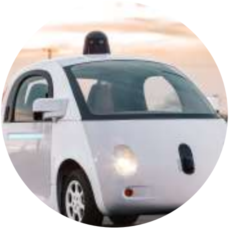
Motonobu Kanagawa
U of Tübingen

Sorawit Saengkyongam
UCL

Sanparith Marukatat
NECTEC

Motivation

Recommendation

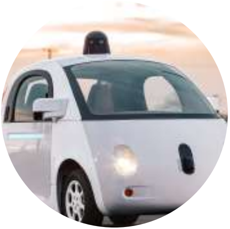


Autonomous Car

Healthcare

Motivation

Recommendation

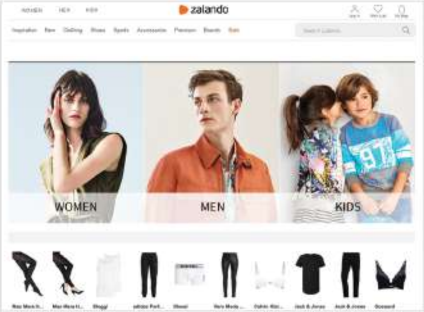


Autonomous Car

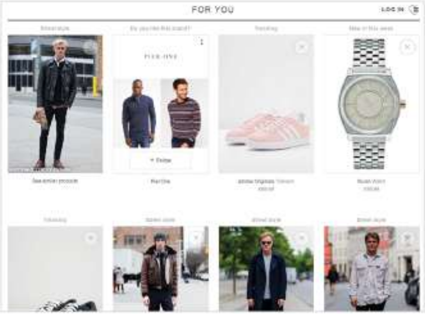
Healthcare

Goal: Identify the best (causal) policy.

Personalization



FIRST VISIT



NEXT VISIT

Healthcare

A Causal Policy

- $\mathcal{X}$: **Context**, $\mathcal{T}$: **Treatment**, $\mathcal{Y}$: **Outcome**, π: **Policy**

[0] The term "context" and "covariate" may be used interchangeably.

A Causal Policy

- $\mathcal{X}$: **Context**, $\mathcal{T}$: **Treatment**, $\mathcal{Y}$: **Outcome**, π: **Policy**
- **Ex:** $\mathcal{X} = \{\text{age}, \text{gender}\}$, $\mathcal{T} = \texttt{pills}$, $\mathcal{Y} = \texttt{cholesterol level}$.

[0] The term "context" and "covariate" may be used interchangeably.

A Causal Policy

- $\mathcal{X}$: **Context**, $\mathcal{T}$: **Treatment**, $\mathcal{Y}$: **Outcome**, π: **Policy**
- **Ex:** $\mathcal{X} = \{\text{age}, \text{gender}\}$, $\mathcal{T} = \texttt{pills}$, $\mathcal{Y} = \texttt{cholesterol level}$.
- A context $x \sim \rho$.

[0] The term "context" and "covariate" may be used interchangeably.

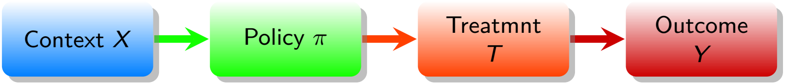
A Causal Policy

- $\mathcal{X}$: **Context**, $\mathcal{T}$: **Treatment**, $\mathcal{Y}$: **Outcome**, π: **Policy**
- **Ex:** $\mathcal{X} = \{\text{age}, \text{gender}\}$, $\mathcal{T} = \texttt{pills}$, $\mathcal{Y} = \texttt{cholesterol level}$.
- A context $x \sim \rho$.
- A treatment $t \sim \pi(t \,|\, x)$ for $(x, t) \in \mathcal{X} \times \mathcal{T}$.

[0] The term "context" and "covariate" may be used interchangeably.

A Causal Policy

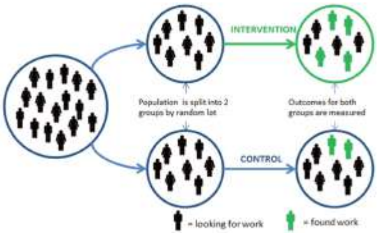
- $\mathcal{X}$: **Context**, $\mathcal{T}$: **Treatment**, $\mathcal{Y}$: **Outcome**, π: **Policy**
- **Ex:** $\mathcal{X} = \{\text{age}, \text{gender}\}$, $\mathcal{T} = \texttt{pills}$, $\mathcal{Y} = \texttt{cholesterol level}$.
- A context $x \sim \rho$.
- A treatment $t \sim \pi(t \,|\, x)$ for $(x, t) \in \mathcal{X} \times \mathcal{T}$.
- An outcome $y \sim \eta(y|x, t)$ for $(x, t, y) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$.



[0] The term "context" and "covariate" may be used interchangeably.

How to Identify Good Policies

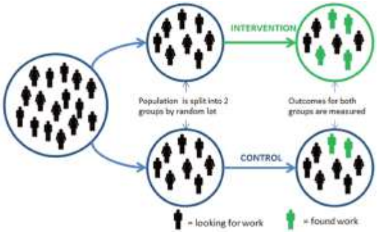
Randomized Exp. (A/B Test)



- ✓ Gold standard in science
- × Expensive, time-consuming, or unethical

How to Identify Good Policies

Randomized Exp. (A/B Test)



- ✓ Gold standard in science
- × Expensive, time-consuming, or unethical

Observational Studies

- ✓ No randomization
- ✓ Cheaper, safer, and more ethical
- × Selection bias

Potential Outcome Framework

- Standard framework in social science, econometric, and healthcare.

Potential Outcome Framework

- Standard framework in social science, econometric, and healthcare.
- Treatment $T \in \{0, 1\}$ and outcome $Y_0, Y_1 \in \mathbb{R}$.
  - $T \in \{\text{placebo}, \text{injection}\}$
  - Y_0 = cholesterol level if T = placebo
  - Y_1 = cholesterol level if T = injection.

Unit	Y_1	Y_0	$Y_1 - Y_0$
A	15	20	-5
B	10	12	-2
C	5	11	-6
D	12	19	-7

(Rubin 2005)

Potential Outcome Framework

- Standard framework in social science, econometric, and healthcare.
- Treatment $T \in \{0, 1\}$ and outcome $Y_0, Y_1 \in \mathbb{R}$.
  - $T \in \{\text{placebo}, \text{injection}\}$
  - $Y_0 =$ cholesterol level if $T =$ placebo
  - $Y_1 =$ cholesterol level if $T =$ injection.

Unit	Y_1	Y_0	$Y_1 - Y_0$
A	15	20	-5
B	10	12	-2
C	5	11	-6
D	12	19	-7

- Individual treatment effect: $\text{ITE}(i) := Y_1(i) - Y_0(i)$

(Rubin 2005)

Potential Outcome Framework

- Standard framework in social science, econometric, and healthcare.
- Treatment $T \in \{0, 1\}$ and outcome $Y_0, Y_1 \in \mathbb{R}$.
  - $T \in \{\text{placebo}, \text{injection}\}$
  - $Y_0 =$ cholesterol level if $T =$ placebo
  - $Y_1 =$ cholesterol level if $T =$ injection.

Unit	Y_1	Y_0	$Y_1 - Y_0$
A	15	-	?
B	-	12	?
C	5	-	?
D	-	19	?

- Individual treatment effect: $\text{ITE}(i) := Y_1(i) - Y_0(i)$
- Fundamental Problem of Causal Inference (FPCI)

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?
- Covariates (X) associated with each unit are available.

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?
- Covariates (X) associated with each unit are available.
- Confounders (Z) affecting both T and Y simultaneously may exist.

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?
- Covariates (X) associated with each unit are available.
- Confounders (Z) affecting both T and Y simultaneously may exist.
- A propensity score:

$$\rho(x) := \mathbb{P}(T = 1 \mid X = x).$$

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?
- Covariates (X) associated with each unit are available.
- Confounders (Z) affecting both T and Y simultaneously may exist.
- A propensity score:

$$\rho(x) := \mathbb{P}(T = 1 \mid X = x).$$

- We observe a dataset

$$\mathcal{D} = \{(x_1, t_1, y_1), (x_2, t_2, y_2), \ldots, (x_n, t_n, y_n)\}$$

where $(x_i, t_i, y_i) :=$ (covariate, received treatment, outcome).

(Rubin 2005)

Rubin's Causal Model

- Causal effect is defined w.r.t. the counterfactual outcomes.
  - What would the value of Y_1 have been had the subject get the injection?
- Covariates (X) associated with each unit are available.
- Confounders (Z) affecting both T and Y simultaneously may exist.
- A propensity score:

$$\rho(x) := \mathbb{P}(T = 1 \,|\, X = x).$$

- We observe a dataset

$$\mathcal{D} = \{(x_1, t_1, y_1), (x_2, t_2, y_2), \ldots, (x_n, t_n, y_n)\}$$

  where $(x_i, t_i, y_i) :=$ (covariate, received treatment, outcome).
- The treatment assignment mechanism is not known.

(Rubin 2005)

Rubin's Causal Model

Main Assumptions

- **Stable unit treatment value assumption (SUTVA)**: The outcome of the ith unit is independent of those of other units and their received treatments.
- **Unconfoundedness/ignorability/exogeneity**

$$Y_0, Y_1 \perp\!\!\!\perp T \mid X$$

- **Treatment positivity**: For all x and t,

$$0 < \mathbb{P}(T = t \mid X = x) < 1.$$

Rubin's Causal Model

Main Assumptions

- **Stable unit treatment value assumption (SUTVA)**: The outcome of the ith unit is independent of those of other units and their received treatments.
- **Unconfoundedness/ignorability/exogeneity**

$$Y_0, Y_1 \perp\!\!\!\perp T \,|\, X$$

- **Treatment positivity**: For all x and t,

$$0 < \mathbb{P}(T = t \,|\, X = x) < 1.$$

Theorem (Propensity Score)

Let $\rho(X) = \mathbb{P}(T = 1 \,|\, X)$ be the propensity score. Suppose that ignorability holds. Then we have

$$Y_0, Y_1 \perp\!\!\!\perp T \,|\, \rho(X).$$

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.
- Our goal is to answer the following counterfactual question:

  "How would the outcomes have changed, if we had switched from the null policy π_0 to the target policy π_1?"

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.
- Our goal is to answer the following counterfactual question:

  "How would the outcomes have changed, if we had switched from the null policy π_0 to the target policy π_1?"
- Let Y_i be the outcome and $Z_i = (X_i, T_i)$ for $i \in \{0, 1\}$.

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.
- Our goal is to answer the following counterfactual question:

  "How would the outcomes have changed, if we had switched from the null policy π_0 to the target policy π_1?"

- Let Y_i be the outcome and $Z_i = (X_i, T_i)$ for $i \in \{0, 1\}$.
- Chernozhukov et al. (2013) defines a **counterfactual distribution**

$$\mathbb{P}_{Y_1} := \int \mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z).$$

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.
- Our goal is to answer the following counterfactual question:

  "How would the outcomes have changed, if we had switched from the null policy π_0 to the target policy π_1?"

- Let Y_i be the outcome and $Z_i = (X_i, T_i)$ for $i \in \{0, 1\}$.
- Chernozhukov et al. (2013) defines a **counterfactual distribution**

$$\mathbb{P}_{Y_1} := \int \mathbb{P}_{Y_0|Z_0}(y|z) \, \mathrm{d}\mathbb{P}_{Z_1}(z).$$

- Under the **main assumptions**, the counterfactual distribution $\mathbb{P}_{Y_1}$ corresponds to the **interventional** distribution $\mathbb{P}^*_{Y_1}$.

Counterfactual Distribution

- π_0: null/logged policy, π_1: target/new policy.
- Our goal is to answer the following counterfactual question:

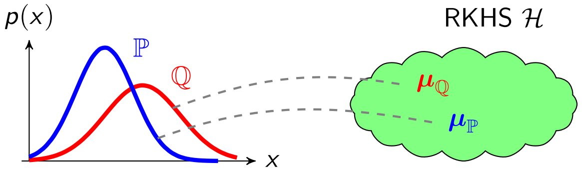
  "How would the outcomes have changed, if we had switched from the null policy π_0 to the target policy π_1?"

- Let Y_i be the outcome and $Z_i = (X_i, T_i)$ for $i \in \{0, 1\}$.
- Chernozhukov et al. (2013) defines a **counterfactual distribution**

$$\mathbb{P}_{Y_1} := \int \mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z).$$

- Under the **main assumptions**, the counterfactual distribution $\mathbb{P}_{Y_1}$ corresponds to the **interventional** distribution $\mathbb{P}^*_{Y_1}$.
- We will construct an estimate for $\mathbb{P}_{Y_1}$ without any sample from it.

Implicit Representation of Distributions

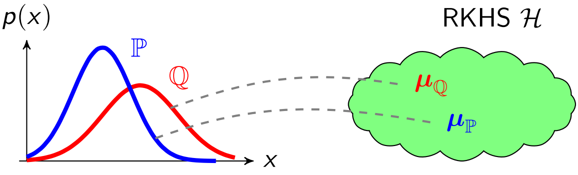


Kernel Mean Embedding (Berlinet and Thomas-Agnan 2004, Smola et al. 2007)

Let $\phi(x) = k(x, \cdot)$ be a canonical feature map from $\mathcal{X}$ into $\mathcal{H}$. A kernel mean embedding (KME) of a distribution $\mathbb{P}$ over $\mathcal{X}$ is defined by

$$\boldsymbol{\mu}_{\mathbb{P}} := \int_{\mathcal{X}} \phi(x)\, \mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}} k(x, \cdot)\, \mathrm{d}\mathbb{P}(x).$$

Implicit Representation of Distributions



Kernel Mean Embedding (Berlinet and Thomas-Agnan 2004, Smola et al. 2007)

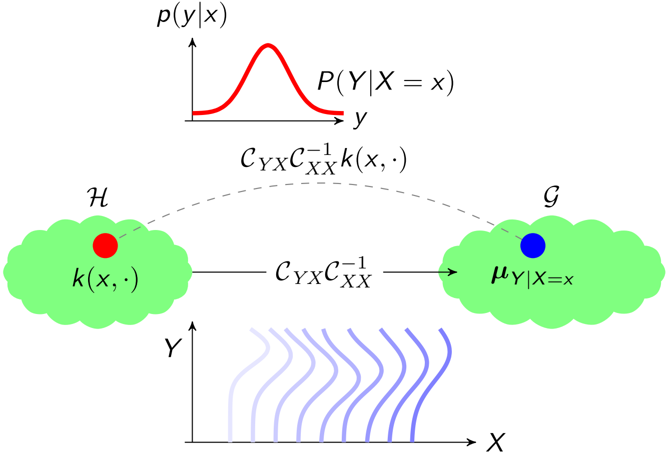
Let $\phi(x) = k(x, \cdot)$ be a canonical feature map from $\mathcal{X}$ into $\mathcal{H}$. A kernel mean embedding (KME) of a distribution $\mathbb{P}$ over $\mathcal{X}$ is defined by

$$\boldsymbol{\mu}_{\mathbb{P}} := \int_{\mathcal{X}} \phi(x)\,\mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}} k(x, \cdot)\,\mathrm{d}\mathbb{P}(x).$$

The embedding $\boldsymbol{\mu}_{\mathbb{P}}$ is well-defined if

1. the kernel k is measurable and
2. the kernel is bounded, i.e., $k(x, x) < \infty$ for all $x \in \mathcal{X}$.

Embedding of Conditional Distributions



The conditional mean embedding of $\mathbb{P}(Y \mid X)$ can be defined as

$$\mathcal{U}_{Y|X} : \mathcal{H} \to \mathcal{G}, \qquad \mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

Counterfactual Mean Embedding

- Recall that we have π_0: null/logged policy, π_1: target/new policy.

Counterfactual Mean Embedding

- Recall that we have π_0: null/logged policy, π_1: target/new policy.
- An embedding of $\mathbb{P}_{Y_1} = \int \mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z)$ can be defined by

$$\boldsymbol{\mu}_{Y_1} := \int \varphi(y)\,\mathrm{d}\mathbb{P}_{Y_1}(y) = \iint \varphi(y)\,\mathrm{d}\mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z) = \mathcal{C}_{Y_0 Z_0}\mathcal{C}_{Z_0}^{-1}\boldsymbol{\mu}_{Z_1}.$$

Counterfactual Mean Embedding

- Recall that we have π_0: null/logged policy, π_1: target/new policy.
- An embedding of $\mathbb{P}_{Y_1} = \int \mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z)$ can be defined by

$$\boldsymbol{\mu}_{Y_1} := \int \varphi(y)\,\mathrm{d}\mathbb{P}_{Y_1}(y) = \iint \varphi(y)\,\mathrm{d}\mathbb{P}_{Y_0|Z_0}(y|z)\,\mathrm{d}\mathbb{P}_{Z_1}(z) = \mathcal{C}_{Y_0 Z_0}\mathcal{C}_{Z_0}^{-1}\boldsymbol{\mu}_{Z_1}.$$

Theorem (causal interpretation)

Suppose that exogeneity holds, i.e., $Y_0, Y_1 \perp\!\!\!\perp T|X$ almost surely for X and that common support assumption holds. Then,

$$\boldsymbol{\mu}_{Y_1} = \boldsymbol{\mu}^*_{Y_1},$$

where $\boldsymbol{\mu}^*_{Y_1}$ denotes an RKHS embedding of the ***interventional distribution*** $\mathbb{P}^*_{Y_1}$.

Counterfactual Mean Embedding

Proposition (empirical estimate)

Given samples $(z_1, y_1), \ldots, (z_n, y_n)$ *from* $\mathbb{P}_{Y_0 Z_0}(z, y)$ *and* $z'_1, \ldots, z'_m$ *from* $\mathbb{P}_{Z_1}(z)$.

- $\Psi = [\varphi(y_1), \ldots, \varphi(y_n)]^\top$
- $\mathbf{K}_{ij} = k(z_i, z_j), \qquad \mathbf{L}_{ij} = k(z_i, z'_j)$
- $\mathbf{1}_n = (1/m, \ldots, 1/m)^\top$

$$\hat{\boldsymbol{\mu}}_{Y_1} = \widehat{\mathcal{C}}_{Y_0 Z_0}(\widehat{\mathcal{C}}_{Z_0} + \varepsilon\mathcal{I})^{-1}\hat{\boldsymbol{\mu}}_{Z_1} = \Psi(\mathbf{K} + n\varepsilon\mathbf{I})^{-1}\mathbf{L}\mathbf{1}_n = \sum_{i=1}^{n} \beta_i \varphi(y_i).$$

Counterfactual Mean Embedding

Proposition (empirical estimate)

Given samples $(z_1, y_1), \ldots, (z_n, y_n)$ *from* $\mathbb{P}_{Y_0 Z_0}(z, y)$ *and* $z'_1, \ldots, z'_m$ *from* $\mathbb{P}_{Z_1}(z)$.

- $\Psi = [\varphi(y_1), \ldots, \varphi(y_n)]^\top$
- $\mathbf{K}_{ij} = k(z_i, z_j), \qquad \mathbf{L}_{ij} = k(z_i, z'_j)$
- $\mathbf{1}_n = (1/m, \ldots, 1/m)^\top$

$$\hat{\boldsymbol{\mu}}_{Y_1} = \widehat{\mathcal{C}}_{Y_0 Z_0}(\widehat{\mathcal{C}}_{Z_0} + \varepsilon \mathcal{I})^{-1} \hat{\boldsymbol{\mu}}_{Z_1} = \Psi(\mathbf{K} + n\varepsilon \mathbf{I})^{-1} \mathbf{L} \mathbf{1}_n = \sum_{i=1}^{n} \beta_i \varphi(y_i).$$

Theorem (uniform convergence)

Under some technical assumptions, if ε_n *decays to zero sufficiently slowly as* $n \to \infty$ *and* $\lim_{n\to\infty} \|\hat{\boldsymbol{\mu}}_{Z_1} - \boldsymbol{\mu}_{Z_1}\|_{\mathcal{H}} = 0$, *we have that, as* $n \to \infty$,

$$\left\|\hat{\boldsymbol{\mu}}_{Y_1} - \boldsymbol{\mu}_{Y_1}\right\|_{\mathcal{G}} \xrightarrow{p} 0.$$

Convergence Rate

Theorem

Let $g := \mathrm{d}\mathbb{P}_{Z_1} / \mathrm{d}\mathbb{P}_{Z_0}$ *and* $\theta(z, \tilde{z}) := \mathbb{E}[\ell(Y_0, \tilde{Y}_0) | Z_0 = z, \tilde{Z}_0 = \tilde{z}]$*. Assume that*

- $g \in \mathrm{Range}(T^{\alpha})$ *for* $0 < \alpha \leq 1$ *and that*
- $\theta \in \mathrm{Range}((T \otimes T)^{\beta})$ *for* $0 < \beta \leq 1$.

Then for $\varepsilon_n = cn^{-1/(1+\beta+\max(1-\alpha,\alpha))}$ *with* $c > 0$ *being arbitrary but independent of* n*, we have*

$$\left\| \widehat{\mathcal{C}}_{Y_0 Z_0} (\widehat{\mathcal{C}}_{Z_0} + \varepsilon_n I)^{-1} \hat{\mu}_{Z_1} - \mu_{Y_1} \right\|_{\mathcal{F}} = O_p \left(n^{-(\alpha+\beta)/(2(1+\beta+\max(1-\alpha,\alpha)))} \right)$$

as $n \to \infty$.

Convergence Rate

Theorem

Let $g := \mathrm{d}\mathbb{P}_{Z_1} / \mathrm{d}\mathbb{P}_{Z_0}$ *and* $\theta(z, \tilde{z}) := \mathbb{E}[\ell(Y_0, \tilde{Y}_0) | Z_0 = z, \tilde{Z}_0 = \tilde{z}]$*. Assume that*

- $g \in \mathrm{Range}(T^{\alpha})$ *for* $0 < \alpha \leq 1$ *and that*
- $\theta \in \mathrm{Range}((T \otimes T)^{\beta})$ *for* $0 < \beta \leq 1$.

Then for $\varepsilon_n = cn^{-1/(1+\beta+\max(1-\alpha,\alpha))}$ *with* $c > 0$ *being arbitrary but independent of* n*, we have*

$$\left\| \widehat{\mathcal{C}}_{Y_0 Z_0} (\widehat{\mathcal{C}}_{Z_0} + \varepsilon_n I)^{-1} \hat{\mu}_{Z_1} - \mu_{Y_1} \right\|_{\mathcal{F}} = O_p \left(n^{-(\alpha+\beta)/(2(1+\beta+\max(1-\alpha,\alpha)))} \right)$$

as $n \to \infty$.

Remark:

- α controls the overlapping between $\mathbb{P}_{Z_1}$ and $\mathbb{P}_{Z_0}$.
- β controls the smoothness of $\mathbb{P}_{Y_0|Z_0}(y|z)$.
- Our estimator has a "doubly-robust"-like property.

Policy Evaluation

- Consider a recommendation platform:
  - **Context:** User information $x \in \mathcal{X}$
  - **Treatment:** Recommendation policy $t \sim \pi(t|x)$
  - **Outcome:** Reward $y = \delta(x, t)$

Policy Evaluation

- Consider a recommendation platform:
  - **Context:** User information $x \in \mathcal{X}$
  - **Treatment:** Recommendation policy $t \sim \pi(t|x)$
  - **Outcome:** Reward $y = \delta(x, t)$
- Given the **logged data** from an initial policy π_0 and target policy π_1:

$$\mathcal{D}_0 = \{(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)\}, \quad \mathcal{D}_1 = \{(x_1^*, t_1^*), \dots, (x_m^*, t_m^*)\}$$

Policy Evaluation

- Consider a recommendation platform:
  - **Context:** User information $x \in \mathcal{X}$
  - **Treatment:** Recommendation policy $t \sim \pi(t|x)$
  - **Outcome:** Reward $y = \delta(x, t)$
- Given the **logged data** from an initial policy π_0 and target policy π_1:

$$\mathcal{D}_0 = \{(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)\}, \quad \mathcal{D}_1 = \{(x_1^*, t_1^*), \dots, (x_m^*, t_m^*)\}$$

- Assume that $\mathbb{P}_0(y \,|\, x', t') = \mathbb{P}_1(y \,|\, x', t')$. Then, we have

$$\mathbb{P}_1(y) = \int \mathbb{P}_1(y \,|\, x^*, t^*) \, \mathrm{d}\mathbb{P}_1(x^*, t^*) = \int \mathbb{P}_0(y \,|\, x, t) \, \mathrm{d}\mathbb{P}_1(x, t)$$

Policy Evaluation

- Consider a recommendation platform:
  - **Context:** User information $x \in \mathcal{X}$
  - **Treatment:** Recommendation policy $t \sim \pi(t|x)$
  - **Outcome:** Reward $y = \delta(x, t)$
- Given the **logged data** from an initial policy π_0 and target policy π_1:

$$\mathcal{D}_0 = \{(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)\}, \quad \mathcal{D}_1 = \{(x_1^*, t_1^*), \dots, (x_m^*, t_m^*)\}$$

- Assume that $\mathbb{P}_0(y \,|\, x', t') = \mathbb{P}_1(y \,|\, x', t')$. Then, we have

$$\mathbb{P}_1(y) = \int \mathbb{P}_1(y \,|\, x^*, t^*) \, \mathrm{d}\mathbb{P}_1(x^*, t^*) = \int \mathbb{P}_0(y \,|\, x, t) \, \mathrm{d}\mathbb{P}_1(x, t)$$

- $\mathbb{P}_1(y)$ is a **counterfactual reward distribution** under the new policy π_1.

Policy Evaluation

- Consider a recommendation platform:
  - **Context:** User information $x \in \mathcal{X}$
  - **Treatment:** Recommendation policy $t \sim \pi(t|x)$
  - **Outcome:** Reward $y = \delta(x, t)$
- Given the **logged data** from an initial policy π_0 and target policy π_1:

$$\mathcal{D}_0 = \{(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)\}, \quad \mathcal{D}_1 = \{(x_1^*, t_1^*), \dots, (x_m^*, t_m^*)\}$$

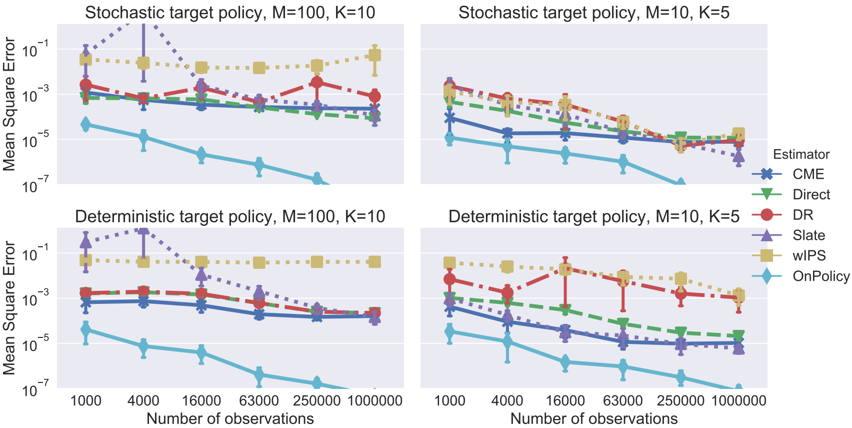
- Assume that $\mathbb{P}_0(y \,|\, x', t') = \mathbb{P}_1(y \,|\, x', t')$. Then, we have

$$\mathbb{P}_1(y) = \int \mathbb{P}_1(y \,|\, x^*, t^*) \,\mathrm{d}\mathbb{P}_1(x^*, t^*) = \int \mathbb{P}_0(y \,|\, x, t) \,\mathrm{d}\mathbb{P}_1(x, t)$$

- $\mathbb{P}_1(y)$ is a **counterfactual reward distribution** under the new policy π_1.
- Let $Z_0 = (X, T)$ and $Z_1 = (X^*, T^*)$.

$$\boldsymbol{\mu}_{\mathbb{P}_1(y)} = \mathcal{C}_{Y_0 Z_0}(\mathcal{C}_{Z_0 Z_0} + \varepsilon\mathcal{I})^{-1}\boldsymbol{\mu}_{Z_1}$$

Experimental Results



Dataset: Microsoft Learning to Rank Challenge dataset (MSLR-WEB30K)

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$
\begin{aligned}
J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)].
\end{aligned}
$$

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\ \nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)]. \end{aligned}$$

- The gradient $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})$ can be directly estimated by CME.

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\ \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)]. \end{aligned}$$

- The gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ can be directly estimated by CME.
- Several disciplines that make use of the **observational studies** will benefit from this work.
  - Social science, econometric, healthcare, finance, etc.

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\ \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)]. \end{aligned}$$

- The gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ can be directly estimated by CME.
- Several disciplines that make use of the **observational studies** will benefit from this work.
  - Social science, econometric, healthcare, finance, etc.
- Include **experimental data** to improve the policy.

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$
\begin{aligned}
J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)].
\end{aligned}
$$

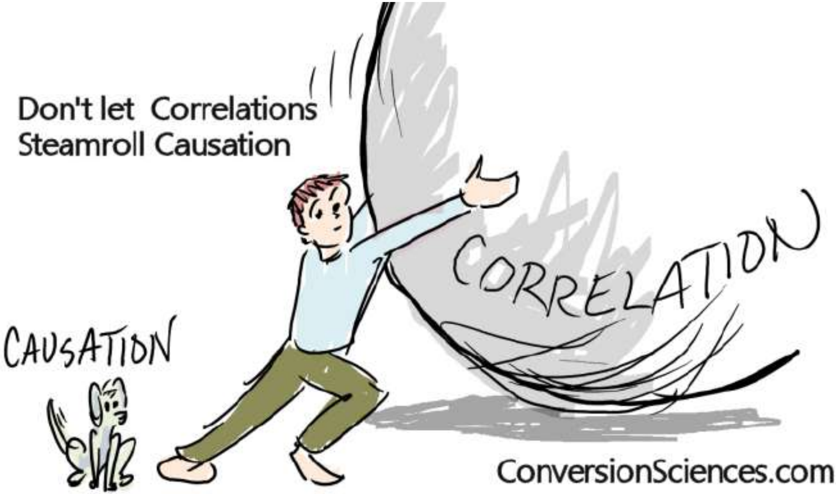
- The gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ can be directly estimated by CME.
- Several disciplines that make use of the **observational studies** will benefit from this work.
  - Social science, econometric, healthcare, finance, etc.
- Include **experimental data** to improve the policy.
- Incorporate multiple sets of observational data obtained from different policies.

Discussion

- In policy learning, given a policy $\pi_{\boldsymbol{\theta}}$, the objective and its gradient are

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \mathbb{E}_{x\sim\rho_X}\mathbb{E}_{t\sim\pi_\theta(t|x)}\mathbb{E}_{y\sim\eta(y|x,t)}\left[\delta(x,t,y)\right] \\ \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\delta(x,t,y)\nabla_{\boldsymbol{\theta}}\log\pi(t|x)]. \end{aligned}$$

- The gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ can be directly estimated by CME.
- Several disciplines that make use of the **observational studies** will benefit from this work.
  - Social science, econometric, healthcare, finance, etc.
- Include **experimental data** to improve the policy.
- Incorporate multiple sets of observational data obtained from different policies.
- Our problem is related to (batch) reinforcement learning, policy gradient methods, and contextual bandit in machine learning.


Don't let Correlations
Steamroll Causation
CORRELATION
CAUSATION
ConversionSciences.com

Contact

Location	Max Planck Campus Tübingen
Website	http://krikamol.org
Email	krikamol@tuebingen.mpg.de
Publication	http://krikamol.org/research/pubs.htm

References I

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.