Statistical Decision Theory with Counterfactual Loss

Benedikt Koch and Kosuke Imai

Department of Statistics, Harvard University

ACIC 2025

Overview

Goal: Evaluate the quality of decisions.

Overview

Goal: Evaluate the quality of decisions.

- Classical decision theory:
  - Evaluates based on *observed outcomes*.
  - Did the decision yield a successful outcome?

Overview

Goal: Evaluate the quality of decisions.

- Classical decision theory:
  - Evaluates based on *observed outcomes*.
  - Did the decision yield a successful outcome?
- This talk:
  - What happens if we use *all potential outcomes*?
  - Would a different decision have produced the same outcome? If so, would it have been preferable?

Overview

Goal: Evaluate the quality of decisions.

- Classical decision theory:
  - Evaluates based on *observed outcomes*.
  - Did the decision yield a successful outcome?
- This talk:
  - What happens if we use *all potential outcomes*?
  - Would a different decision have produced the same outcome? If so, would it have been preferable?

Contribution: Extend classical decision theory for treatment choice to counterfactual losses.

Statistical Decision Theory

Wald 1950: Decision-making as a game against nature.

1. Nature picks an unknown state θ,
2. Decision-maker chooses action $D = d$,
3. A **loss** $\ell(d, \theta)$ quantifies the cost of choosing d under θ.

Statistical Decision Theory

Wald 1950: Decision-making as a game against nature.

1. Nature picks an unknown state θ,
2. Decision-maker chooses action $D = d$,
3. A **loss** $\ell(d, \theta)$ quantifies the cost of choosing d under θ.

Given covariates $\boldsymbol{X}$, construct a decision rule $D = \pi(\boldsymbol{X})$.

Measure performance with **risk**,

$$R(\pi; \theta, \ell) = \mathbb{E}_\theta \left[\ell(\pi(\boldsymbol{X}), \theta) \right].$$

Treatment Choice

Manski [2000; 2004; 2011]: Statistical decision theory for treatment choice.

Idea:

- Choose treatment $D = d$ to minimize loss based on outcome Y.
- Loss depends on potential outcome $Y(d)$, i.e., $\ell(d, Y(d))$.

Treatment Choice

Manski [2000; 2004; 2011]: Statistical decision theory for treatment choice.

Idea:

- Choose treatment $D = d$ to minimize loss based on outcome Y.
- Loss depends on potential outcome $Y(d)$, i.e., $\ell(d, Y(d))$.

Given covariates $\boldsymbol{X}$, use a treatment rule $D = \pi(\boldsymbol{X})$.
Evaluate risk

$$R(\pi; \ell) = \mathbb{E}\left[\ell(\pi(\boldsymbol{X}), Y(\pi(\boldsymbol{X})))\right]$$

Limitation: Loss only depends on the treated potential outcome.

Trichotomous Decision

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment (more invasive)
- $D = 2$: Experimental treatment (most invasive)

c_D cost of treatment D

Trichotomous Decision

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment (more invasive)
- $D = 2$: Experimental treatment (most invasive)

c_D cost of treatment D

Outcome

- $Y = 1$: survival
- $Y = 0$ death

ℓ_y loss under outcome $Y(D) = y$

Trichotomous Decision

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment (more invasive)
- $D = 2$: Experimental treatment (most invasive)

c_D cost of treatment D

Outcome

- $Y = 1$: survival
- $Y = 0$ death

ℓ_y loss under outcome $Y(D) = y$

Standard loss:

$$\ell^{\mathrm{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$$

Trichotomous Decision

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment (more invasive)
- $D = 2$: Experimental treatment (most invasive)

c_D cost of treatment D

Outcome

- $Y = 1$: survival
- $Y = 0$ death

ℓ_y loss under outcome $Y(D) = y$

Standard loss:

$$\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$$

- $(D, Y(D)) = (1, 0) : \ell_0 + c_1$
- $(D, Y(D)) = (2, 1) : \ell_1 + c_2$

Trichotomous Decision II

Standard Loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Clinical & ethical goal: Avoid overtreatment

Trichotomous Decision II

Standard Loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Clinical & ethical goal: Avoid overtreatment

- Prefer least invasive treatment that ensures survival
- Prefer option $k < d$ if $Y(k) = 1$
- r_k regret of overtreating option k

Trichotomous Decision II

Standard Loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Clinical & ethical goal: Avoid overtreatment

- Prefer least invasive treatment that ensures survival
- Prefer option $k < d$ if $Y(k) = 1$
- r_k regret of overtreating option k

Counterfactual loss:

$$\ell^{\text{COF}}(D; Y(0), Y(1), Y(2)) = \ell_{Y(D)} + c_D + \sum_{k<D} r_k Y(k).$$

Trichotomous Decision II

Standard Loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Clinical & ethical goal: Avoid overtreatment

- Prefer least invasive treatment that ensures survival
- Prefer option $k < d$ if $Y(k) = 1$
- r_k regret of overtreating option k

Counterfactual loss:

$$\ell^{\text{COF}}(D; Y(0), Y(1), Y(2)) = \ell_{Y(D)} + c_D + \sum_{k<D} r_k Y(k).$$

$(Y(0), Y(1), Y(2)) = (0, 1, 1)$

- $D = 1 : \ell_1 + c_1$
- $D = 2 : \ell_1 + c_2 + r_1$

Trichotomous Decision II

Standard Loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Clinical & ethical goal: Avoid overtreatment

- Prefer least invasive treatment that ensures survival
- Prefer option $k < d$ if $Y(k) = 1$
- r_k regret of overtreating option k

Counterfactual loss:

$$\ell^{\text{COF}}(D; Y(0), Y(1), Y(2)) = \ell_{Y(D)} + c_D + \sum_{k<D} r_k Y(k).$$

$(Y(0), Y(1), Y(2)) = (0, 1, 1)$

- $D = 1 : \ell_1 + c_1$
- $D = 2 : \ell_1 + c_2 + r_1$

We show:

- For r_k sufficiently large, ℓ^{STD} and ℓ^{COF} yield different treatment preferences.
- No standard loss that can take these ethical considerations into account.

Setup

Observed data: For each unit $i = 1, \ldots, n$, observe $(\boldsymbol{X}_i, D_i, Y_i)$, where:

- Covariates: $\boldsymbol{X}_i \in \mathcal{X}$
- Decision: $D_i \in \mathcal{D} = \{0, 1, \ldots, K-1\}$
- Outcome: $Y_i \in \mathcal{Y} = \{0, 1, \ldots, M-1\}$
- Potential Outcome under $D = d$: $Y(d) \in \mathcal{Y}$

Setup

Observed data: For each unit $i = 1, \ldots, n$, observe $(\boldsymbol{X}_i, D_i, Y_i)$, where:

- Covariates: $\boldsymbol{X}_i \in \mathcal{X}$
- Decision: $D_i \in \mathcal{D} = \{0, 1, \ldots, K-1\}$
- Outcome: $Y_i \in \mathcal{Y} = \{0, 1, \ldots, M-1\}$
- Potential Outcome under $D = d$: $Y(d) \in \mathcal{Y}$

Aim: Study the quality of a generic decision $D_i^* \in \mathcal{D}$ (think of $D^* = \pi(\boldsymbol{X})$)

Setup

Observed data: For each unit $i = 1, \ldots, n$, observe $(\boldsymbol{X}_i, D_i, Y_i)$, where:

- Covariates: $\boldsymbol{X}_i \in \mathcal{X}$
- Decision: $D_i \in \mathcal{D} = \{0, 1, \ldots, K-1\}$
- Outcome: $Y_i \in \mathcal{Y} = \{0, 1, \ldots, M-1\}$
- Potential Outcome under $D = d$: $Y(d) \in \mathcal{Y}$

Aim: Study the quality of a generic decision $D_i^* \in \mathcal{D}$ (think of $D^* = \pi(\boldsymbol{X})$)

Assumptions:

- **IID Sampling:** $\{Y_i, D_i, D_i^*, \boldsymbol{X}_i\}$ are IID
- **Consistency:** $Y_i = Y_i(D_i)$, and if $D_i^* = D_i$, then $Y_i(D_i^*) = Y_i(D_i)$
- **Strong Ignorability:**
  - *Unconfoundedness:* $D_i \perp\!\!\!\perp (D_i^*, \{Y_i(d)\}_{d \in \mathcal{D}}) \mid \boldsymbol{X}_i$
  - *Overlap:* $\exists \eta > 0 : \eta < \Pr(D_i = d \mid \boldsymbol{X}_i) < 1 - \eta$, for all $d \in \mathcal{D}$

Counterfactual Loss and Risk

Counterfactual loss: $\ell : \mathcal{D} \times \mathcal{Y}^K \times \mathcal{X} \to \mathbb{R}$, i.e., $\ell(d; y_1, \ldots, y_k, \boldsymbol{x})$.
Loss of choosing $D^* = d$ given

- Potential outcomes: $(Y(0), \ldots, Y(K-1)) = (y_0, \ldots, y_{K-1})$
- Covariates: $\boldsymbol{X} = \boldsymbol{x}$

Counterfactual Loss and Risk

Counterfactual loss: $\ell : \mathcal{D} \times \mathcal{Y}^K \times \mathcal{X} \to \mathbb{R}$, i.e., $\ell(d; y_1, \ldots, y_k, \boldsymbol{x})$.
Loss of choosing $D^* = d$ given

- Potential outcomes: $(Y(0), \ldots, Y(K-1)) = (y_0, \ldots, y_{K-1})$
- Covariates: $\boldsymbol{X} = \boldsymbol{x}$

Definition (Counterfactual Risk and Conditional Counterfactual Risk)

Given counterfactual loss ℓ, the counterfactual risk of decision D^* is:

$$R(D^*; \ell) := \mathbb{E}\left[\ell(D^*; Y(0), \ldots, Y(K-1), \boldsymbol{X})\right]$$

Counterfactual Loss and Risk

Counterfactual loss: $\ell : \mathcal{D} \times \mathcal{Y}^K \times \mathcal{X} \to \mathbb{R}$, i.e., $\ell(d; y_1, \ldots, y_k, \boldsymbol{x})$.
Loss of choosing $D^* = d$ given

- Potential outcomes: $(Y(0), \ldots, Y(K-1)) = (y_0, \ldots, y_{K-1})$
- Covariates: $\boldsymbol{X} = \boldsymbol{x}$

Definition (Counterfactual Risk and Conditional Counterfactual Risk)

Given counterfactual loss ℓ, the counterfactual risk of decision D^* is:

$$R(D^*; \ell) := \mathbb{E}\left[\ell(D^*; Y(0), \ldots, Y(K-1), \boldsymbol{X})\right] = \mathbb{E}\left[R_{\boldsymbol{X}}(D^*; \ell)\right]$$

where the conditional counterfactual risk given $\boldsymbol{X} = \boldsymbol{x}$ is,

$$R_{\boldsymbol{x}}(D^*; \ell) := \sum_{d \in \mathcal{D}} \sum_{\{y_k\}_{k=0}^{K-1} \in \mathcal{Y}^K} \ell(d; y_0, \ldots, y_{K-1}, \boldsymbol{x}) \times \Pr(D^* = d, Y(0) = y_0, \ldots, Y(K-1) = y_{K-1} \mid \boldsymbol{X} = \boldsymbol{x}).$$

Counterfactual Loss and Risk

Counterfactual loss: $\ell : \mathcal{D} \times \mathcal{Y}^K \times \mathcal{X} \to \mathbb{R}$, i.e., $\ell(d; y_1, \ldots, y_k, \boldsymbol{x})$.
Loss of choosing $D^* = d$ given

- Potential outcomes: $(Y(0), \ldots, Y(K-1)) = (y_0, \ldots, y_{K-1})$
- Covariates: $\boldsymbol{X} = \boldsymbol{x}$

Definition (Counterfactual Risk and Conditional Counterfactual Risk)

Given counterfactual loss ℓ, the counterfactual risk of decision D^* is:

$$R(D^*; \ell) := \mathbb{E}\left[\ell(D^*; Y(0), \ldots, Y(K-1), \boldsymbol{X})\right] = \mathbb{E}\left[R_{\boldsymbol{X}}(D^*; \ell)\right]$$

where the conditional counterfactual risk given $\boldsymbol{X} = \boldsymbol{x}$ is,

$$R_{\boldsymbol{x}}(D^*; \ell) := \sum_{d \in \mathcal{D}} \sum_{\{y_k\}_{k=0}^{K-1} \in \mathcal{Y}^K} \ell(d; y_0, \ldots, y_{K-1}, \boldsymbol{x}) \times \Pr(D^* = d, Y(0) = y_0, \ldots, Y(K-1) = y_{K-1} \mid \boldsymbol{X} = \boldsymbol{x}).$$

Problem: $\Pr(D^* = d, Y(0) = y_0, \ldots, Y(K-1) = y_{K-1} \mid \boldsymbol{X} = \boldsymbol{x})$ unidentifiable

Identifiability of Counterfactual Risk

"Definition": A causal parameter is identifiable if it can be expressed as a function of the observables, i.e., $\Pr(D, D^*, Y, \boldsymbol{X})$.

Identifiability of Counterfactual Risk

"Definition": A causal parameter is identifiable if it can be expressed as a function of the observables, i.e., $\Pr(D, D^*, Y, \boldsymbol{X})$.

Focus on $R_{\boldsymbol{X}}$ (equivalent to R). Issue

$$\Pr(D^*, Y(0), \ldots, Y(K-1) \mid \boldsymbol{X})$$

not identifiable.

Identifiability of Counterfactual Risk

"Definition": A causal parameter is identifiable if it can be expressed as a function of the observables, i.e., $\Pr(D, D^*, Y, \boldsymbol{X})$.

Focus on $R_{\boldsymbol{X}}$ (equivalent to R). Issue

$$\Pr(D^*, Y(0), \ldots, Y(K-1) \mid \boldsymbol{X})$$

not identifiable. However, under strong ignorability

$$\Pr(D^* = d, Y(k) = y_k \mid \boldsymbol{X}) = \Pr(D^* = d, Y = y_k \mid D = k, \boldsymbol{X}).$$

Identifiability of Counterfactual Risk

"Definition": A causal parameter is identifiable if it can be expressed as a function of the observables, i.e., $\Pr(D, D^*, Y, \boldsymbol{X})$.

Focus on $R_{\boldsymbol{X}}$ (equivalent to R). Issue

$$\Pr(D^*, Y(0), \ldots, Y(K-1) \mid \boldsymbol{X})$$

not identifiable. However, under strong ignorability

$$\Pr(D^* = d, Y(k) = y_k \mid \boldsymbol{X}) = \Pr(D^* = d, Y = y_k \mid D = k, \boldsymbol{X}).$$

Can we impose structure on ℓ that enables identification?

Additivity

Definition (Additive Counterfactual Loss)

Let $\boldsymbol{y} = (y_0, \ldots, y_{K-1}) \in \mathcal{Y}^K$. Then the additive counterfactual loss is defined as,

$$\ell^{\mathrm{ADD}}(d; \boldsymbol{y}, \boldsymbol{x}) = \omega_d(d, y_d, \boldsymbol{x}) + \sum_{k \in \mathcal{D}, k \neq d} \omega_k(d, y_k, \boldsymbol{x}) + \varpi(\boldsymbol{y}, \boldsymbol{x}).$$

Additivity

Definition (Additive Counterfactual Loss)

Let $\boldsymbol{y} = (y_0, \ldots, y_{K-1}) \in \mathcal{Y}^K$. Then the additive counterfactual loss is defined as,

$$\ell^{\mathrm{ADD}}(d; \boldsymbol{y}, \boldsymbol{x}) = \omega_d(d, y_d, \boldsymbol{x}) + \sum_{k \in \mathcal{D}, k \neq d} \omega_k(d, y_k, \boldsymbol{x}) + \varpi(\boldsymbol{y}, \boldsymbol{x}).$$

- $\omega_d(d, y_d, \boldsymbol{x}) : \mathcal{D} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
  - Factual weight: Contribution of observed outcome $Y(d)$
  - Decision dependent
  - $\omega_d(d, y_d) = \ell_{y_d} + c_d$

Additivity

Definition (Additive Counterfactual Loss)

Let $\boldsymbol{y} = (y_0, \ldots, y_{K-1}) \in \mathcal{Y}^K$. Then the additive counterfactual loss is defined as,

$$\ell^{\text{ADD}}(d; \boldsymbol{y}, \boldsymbol{x}) = \omega_d(d, y_d, \boldsymbol{x}) + \sum_{k \in \mathcal{D}, k \neq d} \omega_k(d, y_k, \boldsymbol{x}) + \varpi(\boldsymbol{y}, \boldsymbol{x}).$$

- $\omega_d(d, y_d, \boldsymbol{x}) : \mathcal{D} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
  - Factual weight: Contribution of observed outcome $Y(d)$
  - Decision dependent
  - $\omega_d(d, y_d) = \ell_{y_d} + c_d$
- $\omega_k(d, y_k, \boldsymbol{x}) : \mathcal{D} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
  - Counterfactual weight: Contribution of unobserved $Y(k)$
  - Decision dependent
  - $\omega_k(d, y_k) = r_k y_k \mathbb{1}\{k < d\}$

Additivity

Definition (Additive Counterfactual Loss)

Let $\boldsymbol{y} = (y_0, \ldots, y_{K-1}) \in \mathcal{Y}^K$. Then the additive counterfactual loss is defined as,

$$\ell^{\mathrm{ADD}}(d; \boldsymbol{y}, \boldsymbol{x}) = \omega_d(d, y_d, \boldsymbol{x}) + \sum_{k \in \mathcal{D}, k \neq d} \omega_k(d, y_k, \boldsymbol{x}) + \varpi(\boldsymbol{y}, \boldsymbol{x}).$$

- $\omega_d(d, y_d, \boldsymbol{x}) : \mathcal{D} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
  - Factual weight: Contribution of observed outcome $Y(d)$
  - Decision dependent
  - $\omega_d(d, y_d) = \ell_{y_d} + c_d$
- $\omega_k(d, y_k, \boldsymbol{x}) : \mathcal{D} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
  - Counterfactual weight: Contribution of unobserved $Y(k)$
  - Decision dependent
  - $\omega_k(d, y_k) = r_k y_k \mathbb{1}\{k < d\}$
- $\varpi(\boldsymbol{y}, \boldsymbol{x}) : \mathcal{Y}^K \times \mathcal{X} \to \mathbb{R}$
  - Intercept term
  - Decision independent
  - $\varpi(\boldsymbol{y}) = 0$

Additivity Implies Identifiability

Theorem (Additivity Implies Identifiability)

Let ℓ^{ADD} be additive. Then,

$$R(D^*; \ell^{\mathrm{ADD}}) = \sum_{d \in \mathcal{D}} \sum_{k \in \mathcal{D}} \sum_{y \in \mathcal{Y}} \mathbb{E}[\omega_k(d, y, \boldsymbol{x}) \Pr(D^* = d, Y(k) = y \mid \boldsymbol{X})] + \mathbb{E}[C(\boldsymbol{X})],$$

where

$$C(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y}^K} \varpi(\boldsymbol{y}, \boldsymbol{x}) \Pr(\boldsymbol{Y}(\mathcal{D}) = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}),$$

with $\boldsymbol{Y}(\mathcal{D}) = (Y(0), \ldots, Y(K-1))$.

Decomposition into identifiable marginal term and unidentifiable term, not depending on D^*.

Thus an additive loss yields an identifiable risk (up to a constant).

Additivity is Necessary and Sufficient

Can a counterfactual risk be identified under a non-additive loss?

Additivity is Necessary and Sufficient

Can a counterfactual risk be identified under a non-additive loss?

Theorem

Under strong ignorability, the counterfactual risk $R(D^*; \ell)$ *is identifiable (up to a constant) if and only if the loss* ℓ *is additive.*

Outlook

In the paper, we further explore:

- Binary outcome
- Connections between loss and principal strata
- When additive counterfactual losses yield different treatment recommendations than standard losses

Outlook

In the paper, we further explore:

- Binary outcome
- Connections between loss and principal strata
- When additive counterfactual losses yield different treatment recommendations than standard losses

Next steps and extensions:

- Incorporating time-dependent decisions and outcomes
- Relaxing strong ignorability
- Continuous outcomes Y

Outlook

In the paper, we further explore:

- Binary outcome
- Connections between loss and principal strata
- When additive counterfactual losses yield different treatment recommendations than standard losses

Next steps and extensions:

- Incorporating time-dependent decisions and outcomes
- Relaxing strong ignorability
- Continuous outcomes Y

Thank you!

Happy to talk counterfactuals: What should I have done?

Scan the QR code to view the paper.

Binary Outcomes

Corollary

Assume $Y \in \{0,1\}$. *Let* ℓ^{ADD} *be additive. Then,*

$$R_{\boldsymbol{x}}(D^*; \ell^{\text{ADD}}) = \sum_{d \in \mathcal{D}} \zeta_d(d, \boldsymbol{x}) \Pr(D^* = d, Y(d) = 1 \mid \boldsymbol{X} = \boldsymbol{x})$$
$$+ \sum_{d \in \mathcal{D}} \sum_{k \in \mathcal{D}, k \neq d} \zeta_k(d, \boldsymbol{x}) \Pr(D^* = d, Y(k) = 1 \mid \boldsymbol{X} = \boldsymbol{x})$$
$$+ \sum_{d \in \mathcal{D}} \xi(d, \boldsymbol{x}) \Pr(D^* = d \mid \boldsymbol{X} = \boldsymbol{x}) + C(\boldsymbol{x}).$$

where $\zeta_k(d, \boldsymbol{x}) = \omega_k(d, 1, \boldsymbol{x}) - \omega_k(d, 0, \boldsymbol{x})$, $\xi(d, \boldsymbol{x}) = \sum_{k \in \mathcal{D}} \omega_k(d, 0, \boldsymbol{x})$, *and*

$$C(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \{0,1\}^K} \varpi(\boldsymbol{y}, \boldsymbol{x}) \Pr(\boldsymbol{Y}(\mathcal{D}) = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}).$$

Decomposition into accuracy, difficulty, decision and unidentifiable constant term.

Binary Outcomes

Corollary

Assume $Y \in \{0,1\}$. Let ℓ^{ADD} be additive. Then,

$$\begin{aligned} R_{\boldsymbol{x}}(D^*; \ell^{\mathrm{ADD}}) = & \sum_{d \in \mathcal{D}} \zeta_d(d, \boldsymbol{x}) \Pr(D^* = d, Y(d) = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \\ & + \sum_{d \in \mathcal{D}} \sum_{k \in \mathcal{D}, k \neq d} \zeta_k(d, \boldsymbol{x}) \Pr(D^* = d, Y(k) = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \\ & + \sum_{d \in \mathcal{D}} \xi(d, \boldsymbol{x}) \Pr(D^* = d \mid \boldsymbol{X} = \boldsymbol{x}) + C(\boldsymbol{x}). \end{aligned}$$

where $\zeta_k(d, \boldsymbol{x}) = \omega_k(d, 1, \boldsymbol{x}) - \omega_k(d, 0, \boldsymbol{x})$, $\xi(d, \boldsymbol{x}) = \sum_{k \in \mathcal{D}} \omega_k(d, 0, \boldsymbol{x})$, *and*

$$C(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \{0,1\}^K} \varpi(\boldsymbol{y}, \boldsymbol{x}) \Pr(\boldsymbol{Y}(\mathcal{D}) = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}).$$

Decomposition into accuracy, difficulty, decision and unidentifiable constant term.
Choosing weights

$$\omega_d(d, 1) \leq \{\omega_{d'}(d, 0)\}_{d' \neq d} \leq 0 \leq \{\omega_{d'}(d, 1)\}_{d' \neq d} \leq \omega_d(d, 0),$$

yields $\zeta_d(d, \boldsymbol{x}) \leq 0 \leq \zeta_k(d, \boldsymbol{x})$ for all $k \neq d$. Implies risk decreases with accuracy and increases with counterfactual regret.

Binary Decisions

Proposition (Additive Counterfactual Risk with Binary Decision)

Suppose that the decision is binary, i.e., $\mathcal{D} = \{0, 1\}$*. For any additive counterfactual loss* $\ell^{\text{ADD}}(d; \boldsymbol{y}, \boldsymbol{x})$*, we can construct a standard loss* $\ell^{\text{STD}}(d, y_d)$ *such that the risk difference* $R(D^*; \ell^{\text{ADD}}) - R(D^*; \ell^{\text{STD}})$ *does not depend on* D^**.*

If decisions are binary, any additive counterfactual risk admits a standard risk that yields identical treatment recommendations.

Binary Decisions

Proposition (Additive Counterfactual Risk with Binary Decision)

Suppose that the decision is binary, i.e., $\mathcal{D} = \{0, 1\}$. *For any additive counterfactual loss* $\ell^{\text{ADD}}(d; \boldsymbol{y}, \boldsymbol{x})$, *we can construct a standard loss* $\ell^{\text{STD}}(d, y_d)$ *such that the risk difference* $R(D^*; \ell^{\text{ADD}}) - R(D^*; \ell^{\text{STD}})$ *does not depend on* D^*.

If decisions are binary, any additive counterfactual risk admits a standard risk that yields identical treatment recommendations.

However,

- ℓ^{STD} has infinitely many additive counterfactual losses ℓ^{ADD}'s with the same treatment recommendations
- Each of them assigns different values to principal strata

Binary Decisions

Proposition (Additive Counterfactual Risk with Binary Decision)

Suppose that the decision is binary, i.e., $\mathcal{D} = \{0, 1\}$*. For any additive counterfactual loss* $\ell^{\mathrm{ADD}}(d; \boldsymbol{y}, \boldsymbol{x})$*, we can construct a standard loss* $\ell^{\mathrm{STD}}(d, y_d)$ *such that the risk difference* $R(D^*; \ell^{\mathrm{ADD}}) - R(D^*; \ell^{\mathrm{STD}})$ *does not depend on* D^**.*

If decisions are binary, any additive counterfactual risk admits a standard risk that yields identical treatment recommendations.

However,

- ℓ^{STD} has infinitely many additive counterfactual losses ℓ^{ADD}'s with the same treatment recommendations
- Each of them assigns different values to principal strata
- Thus ℓ^{STD} has no clear interpretation based on principal strata
- While ℓ^{ADD} does have a clear interpretation based on principal strata

General Decisions

Proposition (Additive Counterfactual Risk with Non-binary Decision)

Assume that the decision is non-binary, i.e., $K = |\mathcal{D}| \geq 3$. *Then, for any additive counterfactual loss with at least one counterfactual weight* $\omega_k(d, y_k, \boldsymbol{x})$ *depending on the decision* $d \in \mathcal{D}$ *and potential outcome* $y_k \in \mathcal{Y}$ *for* $d \neq k$, *there exists no standard loss* $\ell^{\text{STD}}(d; y_d)$ *such that the risk difference* $R(D^*; \ell^{\text{ADD}}) - R(D^*; \ell^{\text{STD}})$ *does not depend on* D^*.

Example: Binary Decisions

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment

c_D cost of treatment D

Example: Binary Decisions

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment

c_D cost of treatment D

Outcome

- $Y = 1$: survival
- $Y = 0$ death

ℓ_y loss under outcome $Y(D) = y$

Standard loss: $\ell^{\text{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Example: Binary Decisions

Physician treating a patient

- $D = 0$: No treatment
- $D = 1$: Standard treatment

c_D cost of treatment D

Outcome

- $Y = 1$: survival
- $Y = 0$ death

ℓ_y loss under outcome $Y(D) = y$

Standard loss: $\ell^{\mathrm{STD}}(D, Y(D)) = \ell_{Y(D)} + c_D$

Counterfactual loss (extended from Ben-Michael, Greiner, et al. 2024):

$$\ell(D; Y(0), Y(1)) = \ell_{Y(D)} + \tilde{\ell}_{Y(1-D)} + c_D$$

$\tilde{\ell}_y$ loss of counterfactual outcome, $\tilde{\ell}_0 < \tilde{\ell}_1$, i.e. loss is greater when the patient survived under the other treatment (missed positive)

Example: Asymmetric Counterfactual Loss

Hippocratic Oath — "Do no harm": Causing harm with treatment is worse than failing to provide treatment.

Example: Asymmetric Counterfactual Loss

Hippocratic Oath — "Do no harm": Causing harm with treatment is worse than failing to provide treatment.

Loss based on Principal Strata (Ben-Michael, Imai, and Jiang 2024):

$$\begin{aligned}\ell(D; Y(0), Y(1)) = {} & (1 - Y(0))Y(1)\ell_D^{\mathrm{R}} + Y(0)(1 - Y(1))\ell_{1-D}^{\mathrm{H}} \\ & + Y(0)Y(1)\ell_1 + (1 - Y(0))(1 - Y(1))\ell_0 + c_D,\end{aligned}$$

Example: Asymmetric Counterfactual Loss

Hippocratic Oath — "Do no harm": Causing harm with treatment is worse than failing to provide treatment.

Loss based on Principal Strata (Ben-Michael, Imai, and Jiang 2024):

$$\begin{aligned}\ell(D; Y(0), Y(1)) = (1 - Y(0))Y(1)\ell_D^{\mathrm{R}} + Y(0)(1 - Y(1))\ell_{1-D}^{\mathrm{H}} \\ + Y(0)Y(1)\ell_1 + (1 - Y(0))(1 - Y(1))\ell_0 + c_D,\end{aligned}$$

Asymmetry in loss:

$$\underbrace{\Delta^{\mathrm{R}} = \ell_0^{\mathrm{R}} - \ell_1^{\mathrm{R}}}_{\substack{\text{Failure to treat} \\ \text{a responder}}} < \underbrace{\Delta^{\mathrm{H}} = \ell_0^{\mathrm{H}} - \ell_1^{\mathrm{H}}}_{\substack{\text{Harming} \\ \text{a patient}}}.$$

Non-additive loss if $\Delta^{\mathrm{R}} \neq \Delta^{\mathrm{H}}$.

References I

Ben-Michael, Eli, D. James Greiner, et al. (Mar. 2024). *Does AI help humans make better decisions? A methodological framework for experimental evaluation*. arXiv:2403.12108 [cs, econ, q-fin, stat]. DOI: 10.48550/arXiv.2403.12108. URL: http://arxiv.org/abs/2403.12108 (visited on 07/03/2024).

Ben-Michael, Eli, Kosuke Imai, and Zhichao Jiang (2024). "Policy Learning with Asymmetric Counterfactual Utilities". In: *Journal of the American Statistical Association* 0.0. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2023.2300507, pp. 1–14. ISSN: 0162-1459. DOI: 10.1080/01621459.2023.2300507. URL: https://doi.org/10.1080/01621459.2023.2300507 (visited on 09/20/2024).

Manski, Charles F (2000). "Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice". In: *Journal of Econometrics* 95.2, pp. 415–442.

— (2004). "Statistical treatment rules for heterogeneous populations". In: *Econometrica* 72.4, pp. 1221–1246.

References II

Manski, Charles F (2011). "Choosing treatment policies under ambiguity". In: *Annu. Rev. Econ.* 3.1, pp. 25–49.

Wald, Abraham (1950). *Statistical decision functions*. Wiley.