From structural to optimal transport counterfactuals

Lucas De Lara

March 8, 2022

Institut de Mathématiques de Toulouse
Artificial and Natural Intelligence Toulouse Institute

Outline

1. Introduction to counterfactual reasoning
2. Structural counterfactuals, revisited
3. Optimal transport counterfactuals
4. Conclusion

Introduction to counterfactual reasoning

What is a counterfactual?

Factual statement: *Bob is a man, he's 190cm tall*

What is a counterfactual?

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 176cm tall*

What is a counterfactual?

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 176cm tall*

Definition

A *counterfactual* is a statement of the form "Had event A occurred then event B would have occurred". It relates an intervention on the state-of-things to its consequences.

Finding true counterfactuals

How can we assess the truth of these statements?

Finding true counterfactuals

How can we assess the truth of these statements?

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been ???cm tall*

Finding true counterfactuals

How can we assess the truth of these statements?

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been ???cm tall*

We need a model to deduce the counterfactual value(s)

Application in explicability

$X \in \mathbb{R}^d$ Features $S \in \mathbb{R}$ Sensitive $h(X, S) \in \mathbb{R}$ Predictor

Application in explicability

$X \in \mathbb{R}^d$ Features $S \in \mathbb{R}$ Sensitive $h(X, S) \in \mathbb{R}$ Predictor

Did the decision $h(x, s)$ depend on the value s of the sensitive variable?

Application in explicability

$X \in \mathbb{R}^d$ Features $S \in \mathbb{R}$ Sensitive $h(X, S) \in \mathbb{R}$ Predictor

Did the decision $h(x, s)$ depend on the value s of the sensitive variable?

Procedure:

1. Compute x' *the* counterfactual value of x for a change $s \mapsto s'$
2. If $h(x, s) \neq h(x', s')$, then $\|x' - x\|$ furnishes an explanation of the disparate treatment underlining the influence of S

1st way: Nearest counterfactual instance

Find the most similar alternative instance

1st way: Nearest counterfactual instance

Find the most similar alternative instance

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 190cm tall*

1st way: Nearest counterfactual instance

Find the most similar alternative instance

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 190cm tall*

$$(x, s) \mapsto (x, s')$$

1st way: Nearest counterfactual instance

Find the most similar alternative instance

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 190cm tall*

$$(x, s) \mapsto (x, s')$$

(Simplicity/Feasibility) Assumption free and easy to compute

(Unfaithful) Implies that gender and height are independent

(Useless) Non explanatory if h is unaware of S.

2nd way: Structural counterfactuals

Deduce the consequences through Pearl's causal modelling

2nd way: Structural counterfactuals

Deduce the consequences through Pearl's causal modelling

U_0, U_1, U_2 Random seeds

Gender $S = G_0(U_0)$

Height $X_1 = G_1(S, U_1)$

Hired $X_2 = G_2(X_1, S, U_2)$

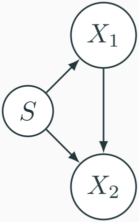

Figure 1: Example of causal graph

2nd way: Structural counterfactuals

Deduce the consequences through Pearl's causal modelling

U_0, U_1, U_2 Random seeds

Gender $S = G_0(U_0)$

Height $X_1 = G_1(S, U_1)$

Hired $X_2 = G_2(X_1, S, U_2)$

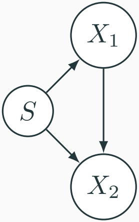

Figure 1: Example of causal graph

(Faithful) Respect structural relationships beyond correlations
(Unfeasible) The causal model is unknown in practice

Towards a 3rd way

Classical approaches tend to be either unfaithful or unfeasible

Towards a 3rd way

Classical approaches tend to be either unfaithful or unfeasible

Counterfactuals must be:

1. Distribution-aware
2. Computationally feasible and assumption-light

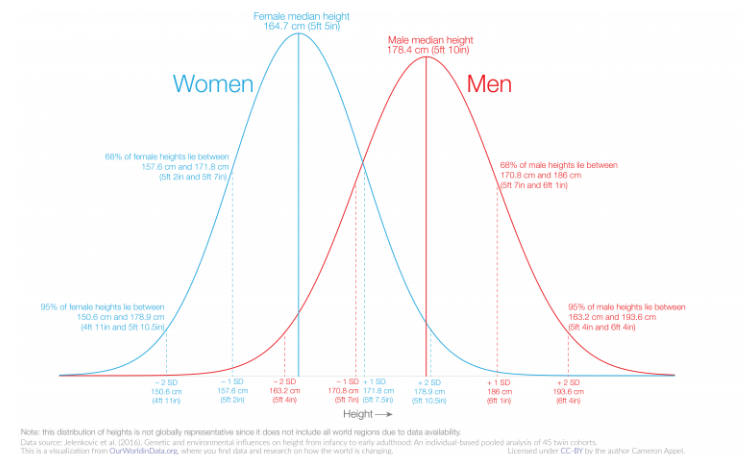

Figure 2: Distribution of female and male height

3rd way: Optimally preserving distributions

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 176cm tall*

3rd way: Optimally preserving distributions

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 176cm tall*

Trading-off causality for correlations

3rd way: Optimally preserving distributions

Factual statement: *Bob is a man, he's 190cm tall*

Counterfactual statement: *Had Bob been a woman, she would have been 176cm tall*

Trading-off causality for correlations

(Faithful) Fits intuition

(Feasible) No assumption on the data-generation process

Structural counterfactuals, revisited

Pearl's causal framework [Pearl, 2009]

Exogenous $U = (U_1, U_2, \ldots)$

Immutable, prior knowledge

Endogenous
$V = (X_1, X_2, \ldots, X_d, S)$

Defined as
$V_i = G_i(V_{\mathrm{Endo}(i)}, U_{\mathrm{Exo}(i)})$

Pearl's causal framework [Pearl, 2009]

Exogenous $U = (U_1, U_2, \ldots)$

Immutable, prior knowledge

Endogenous

$V = (X_1, X_2, \ldots, X_d, S)$

Defined as

$V_i = G_i(V_{\mathrm{Endo}(i)}, U_{\mathrm{Exo}(i)})$

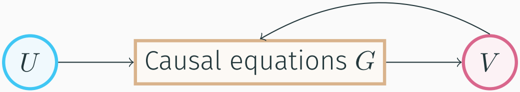

Figure 3: Principle of an SCM

Exogenous $U = (U_1, U_2, \ldots)$

Immutable, prior knowledge

Endogenous $V = (X_1, X_2, \ldots, X_d, S)$

Defined as
$V_i = G_i(V_{\mathrm{Endo}(i)}, U_{\mathrm{Exo}(i)})$

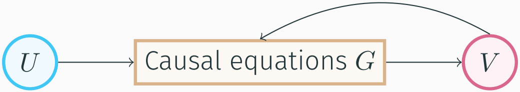

Figure 3: Principle of an SCM

Solvability: There exists a solution map Γ such that $V = \Gamma(U)$

In particular $X = F(S, U_X)$

Do-intervention

Definition of $\mathrm{do}(S = s')$

Forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

Do-intervention

Definition of $\mathrm{do}(S = s')$

Forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

$$X = F(S, U_X) \xrightarrow{\mathrm{do}(S=s')} X_{S=s'} = F(s', U_X)$$

Do-intervention

Definition of $\mathrm{do}(S = s')$

Forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

$$X = F(S, U_X) \xrightarrow{\mathrm{do}(S=s')} X_{S=s'} = F(s', U_X)$$

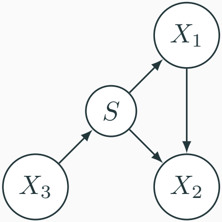

Figure 4: Graph of $\mathcal{M}$

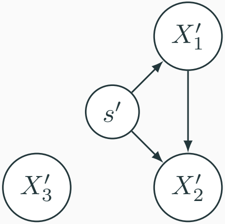

Figure 5: Graph of $\mathcal{M}_{S=s'}$

Pearl's counterfactuals

Counterfactual distribution:

Had S been equal to s' instead of s, X would have follow $\mathcal{L}(X_{S=s'}|S=s)$

Generated by estimating and sampling $\mathcal{L}(U_X|S=s)$

Pearl's counterfactuals

Counterfactual distribution:

Had S been equal to s' instead of s, X would have follow $\mathcal{L}(X_{S=s'}|S=s)$

Generated by estimating and sampling $\mathcal{L}(U_X|S=s)$

Counterfactuals of a single instance x:

Had S been equal to s' instead of s, X would have follow $\mathcal{L}(X_{S=s'}|X=x, S=s)$ instead of δ_x

Generated by estimating and sampling $\mathcal{L}(U_X|X=x, S=s)$

The mass transportation viewpoint

The effect of $\mathrm{do}(S = s' | S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s' | s \rangle} := \mathcal{L}\left((X, X_{S=s'}) | S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

The mass transportation viewpoint

The effect of $\mathrm{do}(S = s'|S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s'|s\rangle} := \mathcal{L}\left((X, X_{S=s'})|S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

Remark:

This coupling admits $\mu_s := \mathcal{L}(X|S = s)$ as first marginal and $\mu_{\langle s'|s\rangle} := \mathcal{L}(X_{S=s'}|S = s)$ as second marginal.

Assumption (RE):
The intervened variable S can be considered a root node of the graph:

$$S \perp\!\!\!\perp U_X$$

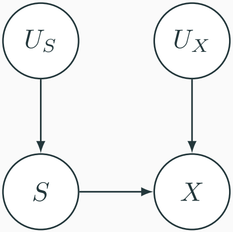

Figure 6: DAG satisfying (RE)

Assumption (RE):
The intervened variable S can be considered a root node of the graph:

$$S \perp\!\!\!\perp U_X$$

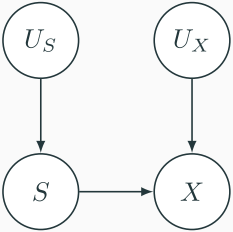

Figure 6: DAG satisfying (RE)

Proposition

If (RE) holds, then

$$\mu_{\langle s' | s \rangle} = \mu_{s'}$$

The exogenous case

Assumption (RE):
The intervened variable S can be considered a root node of the graph:

$$S \perp\!\!\!\perp U_X$$

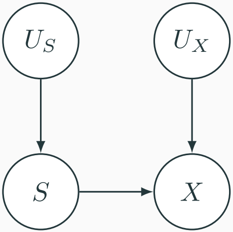

Figure 6: DAG satisfying (RE)

Proposition

If (RE) holds, then

$$\mu_{\langle s' | s \rangle} = \mu_{s'}$$

Consequence: $\pi^*_{\langle s' | s \rangle} \in \Pi(\mu_s, \mu_{s'})$.

The deterministic case

Reminder: $X = F(S, U_X)$

The deterministic case

Reminder: $X = F(S, U_X)$

Assumption (SW): Knowing $S = s$, the model induces a one-to-one relationship between X values and U_X values:

The function $f_s := F(s, \cdot)$ is injective

The deterministic case

Reminder: $X = F(S, U_X)$

Assumption (SW): Knowing $S = s$, the model induces a one-to-one relationship between X values and U_X values:

The function $f_s := F(s, \cdot)$ is injective

Proposition

If (SW) holds, then each instance $x \sim \mu_s$ admits a unique counterfactual counterpart $x' = T^*_{\langle s' | s \rangle}(x)$ where

$$T^*_{\langle s' | s \rangle} := f_{s'} \circ f_s^{-1}|_{\mathcal{X}_s}.$$

In such a scenario, U is unnecessary to compute counterfactuals

An example

Linear additive SCM:

$$
\begin{aligned}
S &= \dots \\
X &= MX + wS + b + U_X
\end{aligned}
$$

An example

Linear additive SCM:

$$
\begin{aligned}
S &= \dots \\
X &= MX + wS + b + U_X
\end{aligned}
$$

Acyclicity implies that $I - M$ is invertible so that

$$X = (I - M)^{-1}(wS + b + U_X) =: F(S, U_X).$$

An example

Linear additive SCM:

$$
\begin{aligned}
S &= \dots \\
X &= MX + wS + b + U_X
\end{aligned}
$$

Acyclicity implies that $I - M$ is invertible so that

$$X = (I - M)^{-1}(wS + b + U_X) =: F(S, U_X).$$

Consequently,

$$T^*_{\langle s' | s \rangle}(x) := x + (I - M)^{-1} w(s' - s).$$

Checkpoint

A counterfactual operation can be characterized by a transport plan between an observable source distribution and a target distribution.

- Under (RE), the target distribution is observable
- Under (SW), the coupling is deterministic (many-to-one)

Checkpoint

A counterfactual operation can be characterized by a transport plan between an observable source distribution and a target distribution.

- Under (RE), the target distribution is observable
- Under (SW), the coupling is deterministic (many-to-one)

	$\neg$(RE)	(RE)
$\neg$(SW)	$\pi^*_{\langle s'\|s\rangle} \in \Pi(\mu_s, \mu_{\langle s'\|s\rangle})$	$\pi^*_{\langle s'\|s\rangle} \in \Pi(\mu_s, \mu_{s'})$
(SW)	${T^*_{\langle s'\|s\rangle}}_\sharp \mu_s = \mu_{\langle s'\|s\rangle}$	${T^*_{\langle s'\|s\rangle}}_\sharp \mu_s = \mu_{s'}$

Shortcomings and limitations

- Finding the causal model (G and U) is too hard in practice (especially when $d \gg 1$)

Shortcomings and limitations

- Finding the causal model (G and U) is too hard in practice (especially when $d \gg 1$)
- A causal-based pipeline would lack efficiency (unrealistic large-scale deployment)

Shortcomings and limitations

- Finding the causal model (G and U) is too hard in practice (especially when $d \gg 1$)
- A causal-based pipeline would lack efficiency (unrealistic large-scale deployment)
- Causal modeling is intrinsically uncertain

Shortcomings and limitations

- Finding the causal model (G and U) is too hard in practice (especially when $d \gg 1$)
- A causal-based pipeline would lack efficiency (unrealistic large-scale deployment)
- Causal modeling is intrinsically uncertain
- Causal counterfactuals may not exist [Bongers et al., 2021]

Optimal transport counterfactuals

Optimal transport

- P, Q probability distributions of $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $\|\cdot - \cdot\|^2$

Optimal transport

- P, Q probability distributions of $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $\|\cdot - \cdot\|^2$

An optimal transport plan $\pi_{P,Q}$ between P and Q w.r.t. cost c is a solution to

$$\min_{\pi \in \Pi(P,Q)} \int \int c(x, x') \mathrm{d}\pi(x, x')$$

Optimal transport

- P, Q probability distributions of $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $\|\cdot - \cdot\|^2$

An optimal transport plan $\pi_{P,Q}$ between P and Q w.r.t. cost c is a solution to

$$\min_{\pi \in \Pi(P,Q)} \int\int c(x, x')\mathrm{d}\pi(x, x')$$

Provides a natural way to create a coupling between two distributions when no canonical choice is available

Surrogate counterfactual model

Under (RE), we know that $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{s'})$... Why not replacing $\pi^*_{\langle s'|s\rangle}$ by an optimal transport plan $\pi_{\langle s'|s\rangle}$ between μ_s and $\mu_{s'}$?

Surrogate counterfactual model

Under (RE), we know that $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{s'})$... Why not replacing $\pi^*_{\langle s'|s\rangle}$ by an optimal transport plan $\pi_{\langle s'|s\rangle}$ between μ_s and $\mu_{s'}$?

Causal counterfactual fairness [Kusner et al., 2017]:

$h(x, s) = h(x', s')$ for any s, s' and (x, x') supported by $\pi^*_{\langle s'|s\rangle}$

Surrogate counterfactual model

Under (RE), we know that $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{s'})$... Why not replacing $\pi^*_{\langle s'|s\rangle}$ by an optimal transport plan $\pi_{\langle s'|s\rangle}$ between μ_s and $\mu_{s'}$?

Causal counterfactual fairness [Kusner et al., 2017]:

$h(x, s) = h(x', s')$ for any s, s' and (x, x') supported by $\pi^*_{\langle s'|s\rangle}$

OT counterfactual fairness:

$h(x, s) = h(x', s')$ for any s, s' and (x, x') supported by $\pi_{\langle s'|s\rangle}$

Theorem [De Lara et al., 2021]

- **Distributions:** μ_s and $\mu_{s'}$ admit densities and have finite second-order moments
- **Causal model:** Both (RE) and (SW) hold
- **Transportation cost:** $c(x, x') = \|x - x'\|^2$

Theorem [De Lara et al., 2021]

- **Distributions:** μ_s and $\mu_{s'}$ admit densities and have finite second-order moments
- **Causal model:** Both (RE) and (SW) hold
- **Transportation cost:** $c(x, x') = \|x - x'\|^2$

$$\pi^*_{\langle s'|s\rangle} = \pi_{\langle s'|s\rangle} \iff f_{s'} \circ f_s^{-1} \text{ is the gradient of a convex function}$$

Theorem [De Lara et al., 2021]

- **Distributions:** μ_s and $\mu_{s'}$ admit densities and have finite second-order moments
- **Causal model:** Both (RE) and (SW) hold
- **Transportation cost:** $c(x, x') = \|x - x'\|^2$

$$\pi^*_{\langle s'|s\rangle} = \pi_{\langle s'|s\rangle} \iff f_{s'} \circ f_s^{-1} \text{ is the gradient of a convex function}$$

The critical assumptions hold for any linear additive model. Recall the example:

Theorem [De Lara et al., 2021]

- **Distributions:** μ_s and $\mu_{s'}$ admit densities and have finite second-order moments
- **Causal model:** Both (RE) and (SW) hold
- **Transportation cost:** $c(x, x') = \|x - x'\|^2$

$$\pi^*_{\langle s'|s\rangle} = \pi_{\langle s'|s\rangle} \iff f_{s'} \circ f_s^{-1} \text{ is the gradient of a convex function}$$

The critical assumptions hold for any linear additive model. Recall the example:

$$f_{s'} \circ f_s^{-1} = x + (I - M)^{-1} w(s' - s).$$

Computational aspects

We don't know μ_s and $\mu_{s'}$ but have access to independent samples.

The OT plan $\pi_{\langle s'|s\rangle}$ must be estimated from data.

Computational aspects

We don't know μ_s and $\mu_{s'}$ but have access to independent samples.

The OT plan $\pi_{\langle s'|s\rangle}$ must be estimated from data.

Exact solver between an n-sample and an m-sample:

- $O((n+m)nm\log(n+m))$ operations
- solution stored as an $n \times m$ matrix

Computational aspects

We don't know μ_s and $\mu_{s'}$ but have access to independent samples.

The OT plan $\pi_{\langle s'|s\rangle}$ must be estimated from data.

Exact solver between an n-sample and an m-sample:

- $O((n+m)nm\log(n+m))$ operations
- solution stored as an $n \times m$ matrix

Growing literature on out-of-samples generalization: plugin estimators, stochastic methods, entropic regularization, generative neural networks...

Practical example

Dataset: Body measurements of 247 men and 260 women.

$X = (\texttt{Weight}, \texttt{Height})$
$S = \texttt{Gender}$

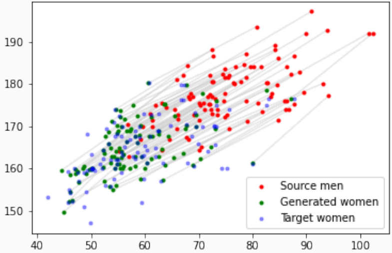

Figure 7: OT intervention

Practical example

Dataset: Body measurements of 247 men and 260 women.

$X = (\mathtt{Weight}, \mathtt{Height})$
$S = \mathtt{Gender}$

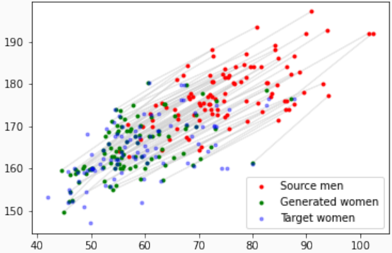

Figure 7: OT intervention

Bob is a 80kg and 190cm man.
Had he been a woman, she would have been 59kg and 177cm.

Remarks

Remarks

Optimal transport trades-off causality for sound correlations, and fits intuition

Remarks

Optimal transport trades-off causality for sound correlations, and fits intuition

Optimal transport solutions are feasible, they can be approximated from data without any assumptions on the data generation process

Remarks

Optimal transport trades-off causality for sound correlations, and fits intuition

Optimal transport solutions are feasible, they can be approximated from data without any assumptions on the data generation process

Optimal transport counterfactuals and structural counterfactuals can be written in a common formalism, making them natural surrogate

Conclusion

Take-away messages

Take-away messages

Counterfactual reasoning

Room for sound correlation-based counterfactuals, between mere translation and causality

Not bound to be either unfaithful or unfeasible

Take-away messages

Counterfactual reasoning

Room for sound correlation-based counterfactuals, between mere translation and causality

Not bound to be either unfaithful or unfeasible

Fairness

Room for individual fairness notions between group fairness and causal fairness

Had my presentation been better, the audience would have asked questions...

References i

Black, E., Yeom, S., and Fredrikson, M. (2020).
Fliptest: Fairness testing via optimal transport.
FAT* '20, page 111–121, New York, NY, USA. Association for Computing Machinery.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021).
Foundations of structural causal models with cycles and latent variables.
The Annals of Statistics, 49(5):2885–2915.

De Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021).
Transport-based counterfactual models.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017).
Counterfactual fairness.
In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.

Pearl, J. (2009).
Causality.
Cambridge university press.

More on the equivalence between SCM and OT counterfactuals

More on the equivalence between SCM and OT counterfactuals

Positive example:

$$
\begin{aligned}
X_1 &= \alpha(S)U_1 + \beta_1(S) \\
X_2 &= -\alpha(S)\ln^2\left(\frac{X_1 - \beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S) \\
S &= U_S \perp\!\!\!\perp (U_1, U_2)
\end{aligned}
$$

More on the equivalence between SCM and OT counterfactuals

Positive example:

$$
\begin{aligned}
X_1 &= \alpha(S)U_1 + \beta_1(S) \\
X_2 &= -\alpha(S)\ln^2\left(\frac{X_1 - \beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S) \\
S &= U_S \perp\!\!\!\perp (U_1, U_2)
\end{aligned}
$$

Negative example:

$$
\begin{aligned}
X_1 &= U_1 \\
X_2 &= SX_1^2 + U_2 \\
S &= U_S \perp\!\!\!\perp (U_1, U_2)
\end{aligned}
$$

Counterfactually fair learning

$$\begin{aligned}\mathcal{R}(\theta) :=& \mathbb{E}[\ell(h_\theta(X, S), Y)] \\ &+ \lambda \sum_{s \in \mathcal{S}} \mathbb{P}(S = s) \sum_{s' \neq s} \mathbb{E}_{\pi_{\langle s' | s \rangle}} \left[\left| h_\theta(X, s) - h_\theta(X', s') \right|^2 \right]\end{aligned}$$

Counterfactually fair learning

$$\begin{aligned}\mathcal{R}(\theta) :=& \mathbb{E}[\ell(h_\theta(X,S),Y)] \\ &+\lambda \sum_{s\in\mathcal{S}} \mathbb{P}(S=s) \sum_{s'\neq s} \mathbb{E}_{\pi_{\langle s'|s\rangle}}\left[\left|h_\theta(X,s)-h_\theta(X',s')\right|^2\right]\end{aligned}$$

Theorem

Under some assumptions (compactness, density, linearity),

$$\mathcal{R}(\theta_n) - \min_{\theta\in\Theta} \mathcal{R}(\theta) \xrightarrow[n\to+\infty]{a.s.} 0.$$

Counterfactually fair learning

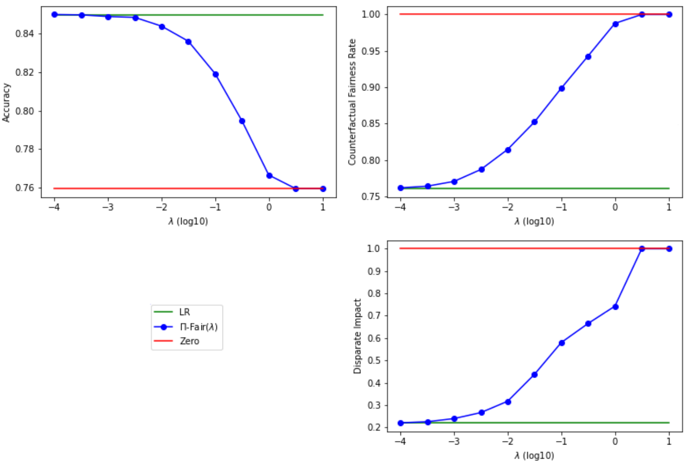

Figure 8: Acc, CFR and DI of the baseline predictors and regularized predictors.