Achieving Counterfactual Fairness in Reinforcement Learning

Jitao Wang[1], Chengchun Shi[2] and Zhenke Wu[1]

[1]University of Michigan, Ann Arbor
[2] London School of Economics and Political Science

Contents

Background and Motivation

Fairness in Decision Making

The growing use of machine learning for automated decision-making has raised concerns about the potential for unfairness in learning algorithms and models.

- University admissions

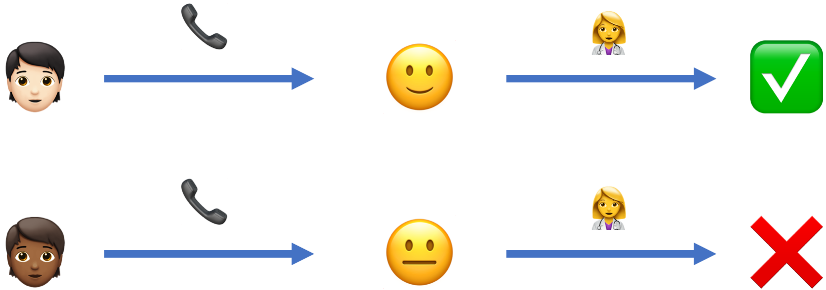
- Hiring

Motivated Data Example

- Prescription Opioid Wellness and Engagement Research in the emergency department (PowerED) study
- Study Population: patients with pain discharged from the emergency department who reported any opioid analgesic (OA) misuse in the past 3 months.
- Treatment options:
  1. a brief interactive voice response (IVR) call (< 5 mins)
  2. a longer motivational IVR call (5 - 10 mins)
  3. a live call from a counselor (20 mins).
- Treatment assignment: Online reinforcement learning (RL) algorithm.
- The goal: minimize OA risk, defined as Current Opioid Misuse Measure (COMM) score
- However, the naive RL algorithm may lead to **unfair** treatment assignment.

Motivated Data Example

The naive RL algorithm may lead to **unfair** treatment assignment.

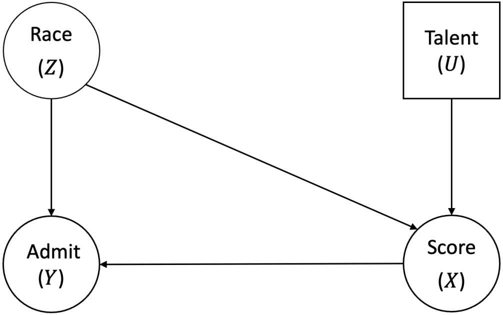


The algorithm always assigns the treatment to those who benefits most, which leads to unfairness.

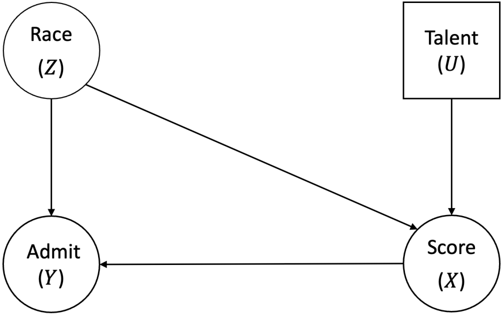
Single-stage fairness

- **Individual fairness**: similar individuals should have similar decisions.
- **Equal opportunity**: equally qualified individuals should have the same decisions.
- **Counterfactual fairness**: an individual should have the same decisons even if he/she belongs to other groups.

Consider university admission example



Single-stage fairness



- **Individual fairness**: Black (white) students with similar SAT scores should have similar probability of admission.
- **Equal opportunity**: Students with similar SAT scores should have similar probability of admission, regardless of their race.
- **Counterfactual fairness**: Students with similar talent should have similar probability of admission, regardless of their race and SAT scores.

Contributions

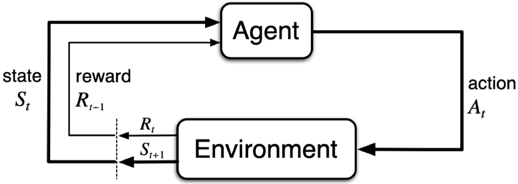
In this work, we

- extend the definition of counterfactual fairness to RL problems;
- propose algorithms to achieve counterfactual fairness in RL;
- demostrate the conditions under which the proposed algorithms can achieve optimal value.

Preliminary

Basics of Reinforcement Learning (RL)

- A *Markov Decision Process* (MDP) consists of:
  - **State** $s \in \mathcal{S}$, information about the environment;
  - **Action** $a \in \mathcal{A}$, the action the agent chooses to take;
  - **Reward** r the immediate reward the agent received from the environment;
  - **Dynamics** of the environment $p_t(s', r|s, a)$, representing the probability of transitioning to state s' and receiving reward r from state s by taking action a at time t.
  - **Policy** $\pi_t(a|s)$, representing the probability of taking action a in state s at time t.



Basics of Reinforcement Learning (RL)

- The goal of RL is to find a policy π that maximizes the sum of discounted rewards:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

  where $\gamma \in [0, 1)$ is the discount factor.
- Value function is defined as

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s\right]$$

- Bellman Optimality Equation:

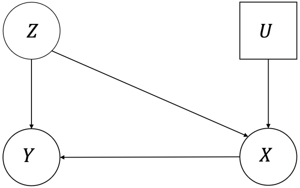
$$V^{\pi^*}(s) = \max_{a} \mathbb{E}_{s',r|s,a}\left[r + \gamma V^{\pi^*}(s')\right]$$

- Optimal policy π^* can be retrived through optimal value function V^{π^*}:

$$\pi^*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_a \mathbb{E}_{s',r|s,a}\left[r + \gamma V^{\pi^*}(s')\right] \\ 0 & \text{otherwise} \end{cases}$$

Basics of Counterfactual Fairness

- Counterfactual fairness was originally proposed in [1].
- Consider a causal graph



Definition 1 (Single-stage counterfactual fairness).

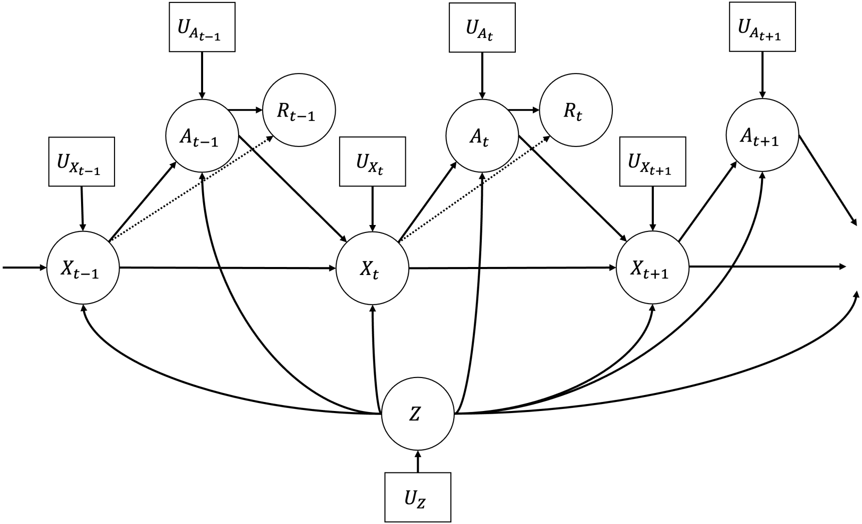
Decision Y is counterfactually fair if under any pair of attributes (x^*, z^*),

$$Y_{z'}(U)|\{X = x^*, Z = z^*\} \overset{d}{=} Y_{z^*}(U)|\{X = x^*, Z = z^*\}$$

for any value z'.

MDP with sensitive attributes

We consider counterfactual fairness in the context of MDP with time invariant sensitive attributes (S-MDP).



Counterfactual Fairness in S-MDP

Define $\bar{x}_t = \{x_0, \ldots, x_t\}, \bar{a}_t = \{a_0, \ldots, a_t\}, \bar{u}X_t = \{u_{X_0}, \ldots, u_{X_t}\}$, then we have

Definition 2 (Counterfactual fairness for policy in S-MDP).

A policy π is counterfactually fair if it satisfies the following condition:

$$A^{\pi}_{t, \boxed{z'}, \bar{a}^*_{t-1}}(\bar{U}_{X_t}) | \{\bar{X}_t = \bar{x}^*_t, \bar{A}_{t-1} = \bar{a}^*_{t-1}, Z = \boxed{z^*}\}$$
$$\stackrel{d}{=} A^{\pi}_{t, \boxed{z^*}, \bar{a}^*_{t-1}}(\bar{U}_{X_t}) | \{\bar{X}_t = \bar{x}^*_t, \bar{A}_{t-1} = \bar{a}^*_{t-1}, Z = \boxed{z^*}\} \quad \text{for } t = 1, \ldots, T.$$

for any z'.

- A person with action trajectory $\bar{a}^*_{t-1}$ will receive the same action A_t, regardless the value of z.
- This is equavalent to say, any person with same action trajectory $\bar{a}^*_{t-1}$ and $\bar{u}^*_{X_t}$ will receive the same action A_t, regardless the value of z.

Intuition & Challenge

Intuition

- Can we remove the information of Z from the state variable X_t so that the value of Z is independent of X_t?
- The modified state variable $\tilde{X}_t$ should only be a function of $\bar{U}_{X_{t-1}}$ and $\bar{A}_{t-1}$.
- Then we can learn the policy that only depends on $\tilde{X}_t$ to achieve counterfactual fairness.

Challenge

- Z not only affects the value X_t directly, but also affects the value of X_{t-1} indirectly through A_{t-1}, X_{t-1}.
- $\bar{U}_{X_t}$ is not observable.
- $\{\tilde{X}_t, A_t, R_t\}$ may form a high-order MDP, leading to suboptimal policy if common RL algorithms are adopted.

Achieving Counterfactual Fairness in S-MDP

General Theorem

Theorem 3 (Achieving Counterfactual Fairness In S-MDP).

Suppose the gradient $\nabla_{\bar{u}_{X_t}} f_{X_t}(z, x_{t-1}, a_{t-1})$ does not involve z for $t = 1, ..., T$, then by applying the following procedure $\mathcal{P}$ on x_t,

$$\tilde{x}_t = \mathcal{P}(x_t, z, \bar{a}_{t-1}) = \sum_{z'} \widehat{x}_t(z', \bar{a}_{t-1})\mathbb{P}(Z = z')$$

where $\widehat{x}_t(z', \bar{a}_{t-1}) = x_t(z, \bar{a}_{t-1}) - \mathbb{E}_n(X_t | Z = z, \bar{A}_{t-1} = \bar{a}_{t-1}) + \mathbb{E}_n(X_t | Z = z', \bar{A}_{t-1} = \bar{a}_{t-1})$, the policy learning algorithms that use the preprocessed experience tuples $\{\tilde{x}_t, a_t, r_t\}$ will be counterfactually fair. Here $x_t \equiv f_{X_t}(z, x_{t-1}, a_{t-1})$.

Key point is that under the condition of Theorem 3, the information of z can be removed through $x_t(z, \bar{a}_{t-1}) - \mathbb{E}_n(X_t | Z = z, \bar{A}_{t-1} = \bar{a}_{t-1}) + \mathbb{E}_n(X_t | Z = z', \bar{A}_{t-1} = \bar{a}_{t-1})$.

Pratical issue: when t is large, some combinations of $\{z, \bar{a}_{t-1}\}$ may not be observed.
We call this procedure **model-free preprocessing** since there is not assumption on models.

S-MDP with Linear model

Corollary 4 (S-MDP with linear model).

A S-MDP with linear model satisfies Theorem 3 if and only if there is no interaction terms (z, x_{t-1}), (z, u_{X_t}) *and* (x_{t-1}, u_{X_t}) *in* $f_{X_t}(z, x_{t-1}, a_{t-1})$

The resulting linear model looks like

$$X_t = \gamma_0 + \gamma_1 Z + \gamma_2 X_{t-1} + \gamma_3 A_{t-1} + \gamma_4 Z A_{t-1} + \gamma_5 X_{t-1} A_{t-1} + \gamma_6 A_{t-1} U_{X_t} + U_{X_t}$$

Given Z and $\bar{a}_{t-1}$, X_t has the following form

$$X_t = b_0^{(t)} + b_1^{(t)} Z + g(\bar{U}_{X_t})$$

where b_1^t can recursively calculated using estimated model parameters $\{\widehat{\gamma}_0, \ldots, \widehat{\gamma}_5\}$.
So we can apply the modifed procedure $\mathcal{P}$ on x_t,

$$\tilde{x}_t = x_t - \widehat{b}_1^{(t)} z + \widehat{b}_1^{(t)} \mathbb{E}(Z)$$

We call this procedure **model-based preprocessing** since it assumes the model is correct.

General S-MDP

Corollary 5 (General S-MDP).

A general S-MDP satisfies Theorem 3 if and only if there is no interaction terms (z, x_{t-1}), (z, u_{X_t}) *and* (x_{t-1}, u_{X_t}) *in* $f_{X_t}(z, x_{t-1}, a_{t-1})$.

The resulting model looks like

$$f_{X_t}(z, x_{t-1}, a_{t-1}) = g_1(z, a_{t-1}) + g_2(x_{t-1}, a_{t-1}) + g_3(u_{X_t}, a_{t-1}),$$

where $g_i, i = 1, 2, 3$ are continuous functions.

Without linearity assumption, we will need to calculate $\mathbb{E}_n(X_t | Z = z, \bar{A}_{t-1} = \bar{a}_{t-1})$ for any t. It is okay for small t, but it will be infeasible for large t.

Possible solutions: 1. use lag-K approximation, or 2. use regression to approximate the conditional expectation.

Optimality in S-MDP with Counterfactual Fairness

General Theorem

Theorem 6 (Optimality in general S-MDP).

For a S-MDP $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{Z}\}$, if the following conditions are satisfied,

$$P(S_{t+1}, R_t | \bar{S}_t, \bar{A}_t, Z) = P(S_{t+1}, R_t | S_t, A_t, Z) \quad \text{for any } t = 1, \ldots, T, \tag{1}$$

$$P(S_{t+1} | S_t, A_t, Z) = P(S_{t+1} | S_t, A_t) \quad \text{for any } t = 1, \ldots, T, \tag{2}$$

$$P(S_0 | Z) = P(S_0) \tag{3}$$

Then there exists some $\pi^{opt} : \mathcal{S} \to \mathcal{A}$ that belong to stationary policies, such that $V(\pi^{opt}, s) = \sup_{\pi \in HR} V(\pi, s)$ for any $s \in \mathcal{S}$ where HR stands for history-dependent policies.

This is a general result for any S-MDP. (1) is the Markov property conditioning on Z. (2) together with (3) require that $S_t \perp Z$ for any $t = 1, \ldots, T$. The processed $\tilde{X}_t$ naturally satisfies the conditions (2) and (3), therefore we can apply this theorem to our setting.

The central point here is that although $\tilde{X}_t$ is Markovian, reward R_t can depend on previous actions $\bar{a}_{t-1}$, making it non-Markovian. Then the optimal policy will not be stationary policy, instead it will be history-dependent.

Optimality in S-MDP with Counterfactual Fairness

Corollary 7 (Optimality in S-MDP with linear model).

The optimal policy trained using the preprocessed experience tuples $\{(\tilde{x}_{it}, a_{it}, r_{it}) : i = 1, \dots, N; t = 1, \dots, T\}$ from a S-MDP with linear model is stationary if and only if there is no interaction terms (z, a_{t-1}) and (x_{t-1}, a_{t-1}) in $f_{X_t}(z, x_{t-1}, a_{t-1})$.

The following linear model satisfies the condition of Corollary 7:

$$
\begin{aligned}
X_t &= \gamma_0 + \gamma_1 Z + \gamma_2 X_{t-1} + \gamma_3 A_{t-1} + \gamma_6 A_{t-1} U_{X_t} + U_{X_t} \\
R_t &= \lambda_0 + \lambda_1 X_t + \lambda_2 Z + \lambda_3 A_t + \lambda_4 X_t Z + \lambda_5 X_t A_t + \lambda_6 Z A_t
\end{aligned}
$$

Still working on general S-MDP.

Simulation

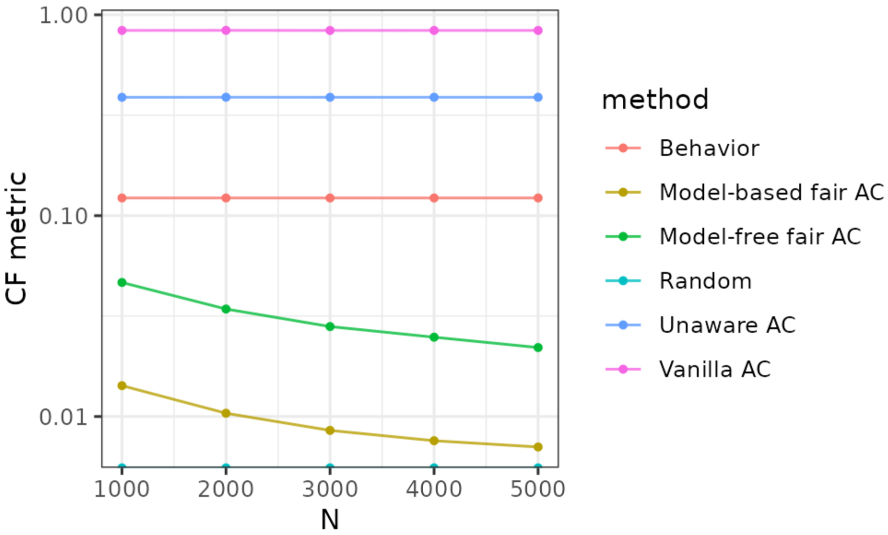
Simulation setup

Data generating process:

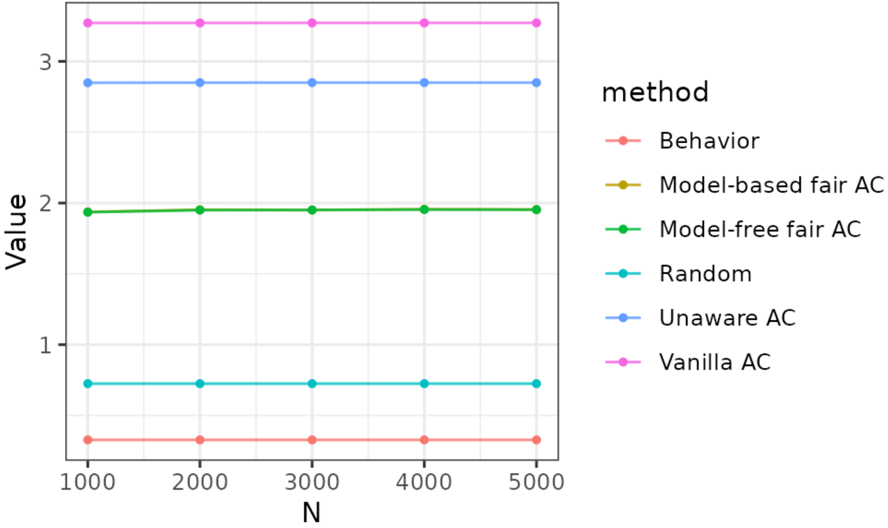
$$
\begin{aligned}
U_{X_t} &\sim \mathcal{N}(0,1)\\
U_{A_t} &\sim \mathcal{U}(0,1)\\
A_t &= I\{U_{A_t} \leq -0.5 + 0.5Z\}\\
Z &\sim \text{Bernoulli}(0.5)\\
X_0 &= -0.7 + 0.8Z + U_{X_0}\\
X_t &= -0.7 + 0.8Z + 0.5X_{t-1} + 0.4A_{t-1} + U_{X_t}\\
R_t &= -0.2 + 0.3X_t + 0.8Z + 0.8A_t - 0.6X_tZ - 0.7X_tA_t - 1.6ZA_t
\end{aligned}
$$

- $N = 1000, 2000, 3000, 4000, 5000$
- $T = 5$
- Use actor critic as policy learning algorithm
- Policies: random, behavior, vanilla actor critic, actor critic with fairness through unawareness, actor critic with model-based preprocessing, actor critic with model-free preprocessing

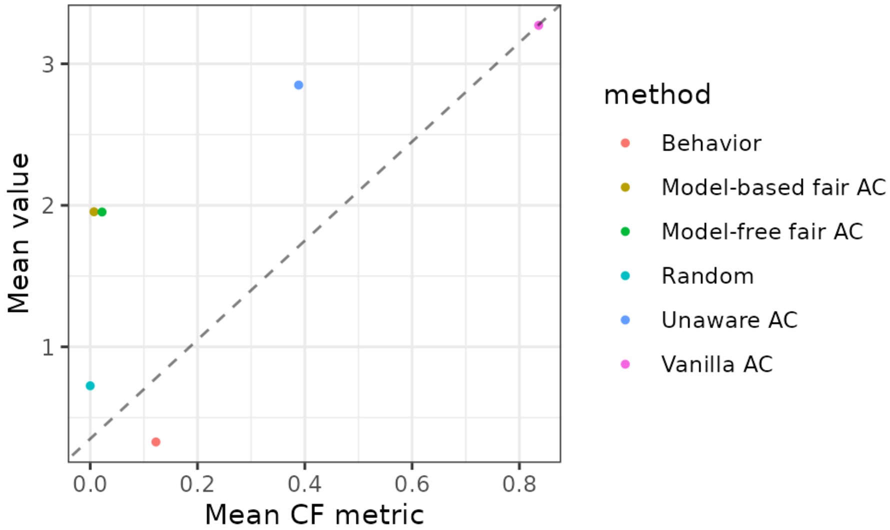
Simulation results (1)



Simulation results (2)



Simulation results (3)



Thank you

[1] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.