Counterfactually Fair Reinforcement Learning via Sequential Data Preprocessing

Jitao Wang[1], Chengchun Shi[2], John D. Piette[1], Joshua R. Loftus[2], Donglin Zeng[1], Zhenke Wu[1]*

[1] University of Michigan [2] London School of Economics and Political Science | jitwang@umich.edu, *zhenkewu@umich.edu

Motivating Example: The PowerED Study

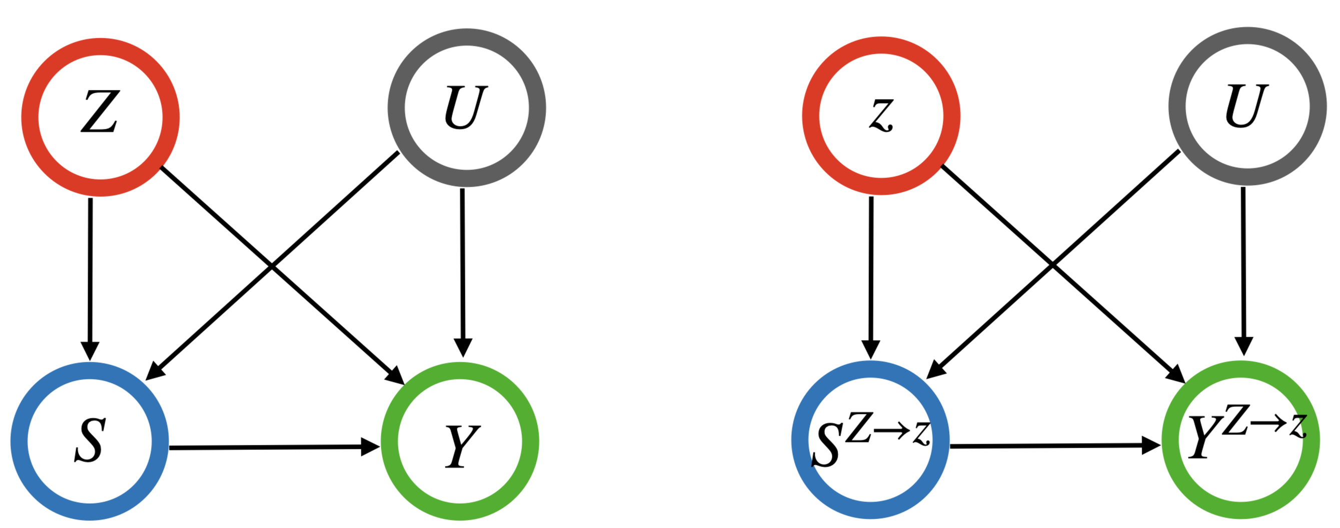
- **Study Goal**: Evaluate whether a 12-week reinforcement learning (RL)-based intervention reduces opioid analgesic (OA) misuse.
- **Treatment:** Each week, an online bandit algorithm assigns patients to one of 1) brief IVR call (<5mins), 2) longer IVR call (5-10 mins), and 3) live call with counselor (~20 mins). Self-reported responses to weekly surveys and baseline information (e.g., COMM score, pain severity) are used as contextual variables.
- **Outcome:** self-reported OA misuse score.
- **Unfairness might arise**: Hispanics may under-report pain levels due to cultural factors, misleading the RL agent to assign less therapist time

Contributions

- Conceptualize counterfactual fairness (CF), a causal based fairness metric, in RL.
- Characterize the class of CF policies and demonstrate the form of the optimal CF policy under stationarity.
- Develop a sequential data preprocessing algorithm for fair policy learning.
- Theoretical guarantees for asymptotic unfairness control and regret bounds.

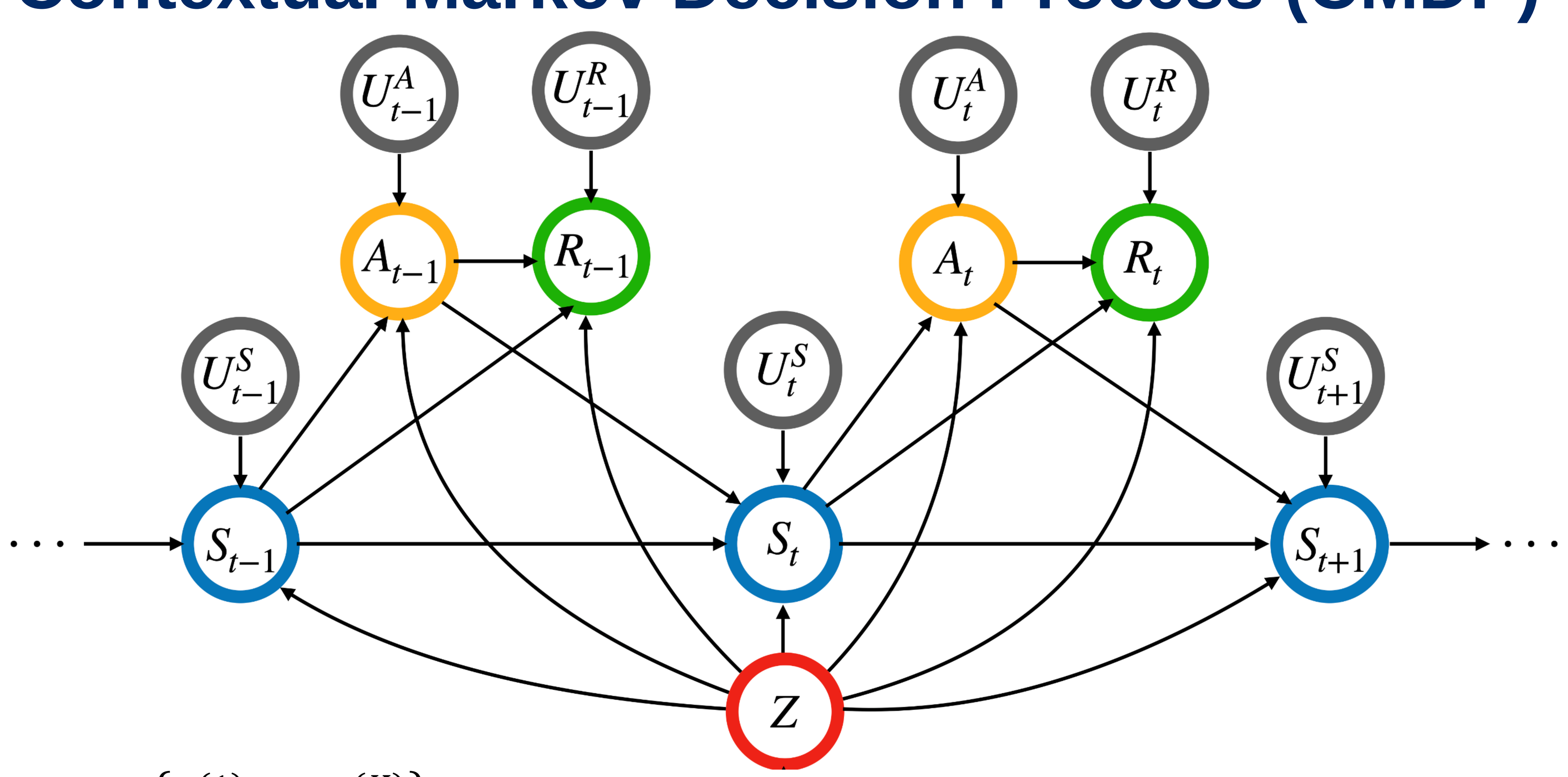
Preliminaries

Counterfactual Fairness (CF)



- **CF:** $P\left(\hat{Y}^{Z\to z}(U)=y \middle| S=s, Z=z\right)=P\left(\hat{Y}^{Z\to z'}(U)=y \middle| S=s, Z=z\right)$

Contextual Markov Decision Process (CMDP)



- $Z \in \{z^{(1)}, \dots, z^{(K)}\}$ is set of all levels of a sensitive attribute.
- S_t = state; A_t = action; R_t = reward; $U^{(\cdot)}$ = exogenous variable,
- History up to time t $H_t = \{Z, \bar{S}_t, \bar{A}_{t-1}, \bar{R}_{t-1}\}$.

Counterfactual Fairness under CMDP

Definition (CF in CMDP). Given an observed trajectory $H_t = h_t = \{z, \bar{a}_{t-1}, \bar{r}_{t-1}, \bar{s}_t\}$, a decision rule π_t is counterfactually fair at time t if it satisfies the following condition:

$$P^{\pi_t}\left(A_t^{Z\leftarrow z'}\left(\bar{U}_t(h_t)\right)=a\right)=P^{\pi_t}\left(A_t^{Z\leftarrow z}\left(\bar{U}_t(h_t)\right)=a\right)$$

for any $z' \in Z$ and $a \in A$ and $\bar{U}_t(\cdot)=\{U_1^S(\cdot), U_1^R(\cdot), \dots, U_{t-1}^S(\cdot), U_{t-1}^R(\cdot), U_t^S(\cdot)\}$.

Candidate	Talent	Gender	Pre-college school level	SAT score
A	**100**	**Female**	**Top**	**1500**
	100	Male	Top	1550
B	**100**	**Male**	**Top**	**1550**
	100	Female	Top	1500

Theorem 1 (Counterfactual augmentation). Given observed history $H_t = h_t$ under CMDPs, π_t satisfies CF if it admits the form $\pi_t(\bar{\mathcal{S}}_t, \bar{\mathcal{R}}_t, \bar{a}_{t-1})$ for any t where

all counterfactual states at time t

$$\mathcal{S}_t = \{S_t^{Z\leftarrow z^{(k)}}(\bar{U}_t(h_t))\}_{k=1,\dots,K} \text{ and } \bar{\mathcal{S}}_t = \{\mathcal{S}_{t'}\}_{t'\le t},$$

all counterfactual rewards at time t

$$\mathcal{R}_t = \{R_t^{Z\leftarrow z^{(k)}}(\bar{U}_t(h_{t+1}))\}_{k=1,\dots,K} \text{ and } \bar{\mathcal{R}}_t = \{\mathcal{R}_{t'}\}_{t'\le t}.$$

Theorem 2. (Stationarity of optimal CF policy) Let HCF denote the class of policies $\pi = \{\pi_t\}_{t\ge1}$ where each π_t maps $(\bar{\mathcal{S}}_t, \bar{\mathcal{R}}_t, \bar{a}_{t-1})$ to a probability mass function of A. Let SCF denote the class of $\pi = \{\pi_t\}_{t\ge1} \in HCF$ for which there exists some function π^* such that $\pi_t(\bar{\mathcal{S}}_t, \bar{\mathcal{R}}_t, \bar{a}_{t-1}) = \pi^*(\bar{\mathcal{S}}_t)$ for any $t \ge 1$ almost surely. Then, under stationary CMDP, there exists some $\pi^{opt} \in SCF$ such that

$$J(\pi^{opt}) = \sup_{\pi\in HCF} J(\pi),$$

where $J(\pi) = E_\pi\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]$ with discount factor $\gamma \in (0,1)$.

Takeaways: Under stationary CMDPs, we only need to focus on stationary policies.

Sequential Preprocessing Algorithm

- **Assumption 1:** For any $t < T$, conditioning on H_t blocks all backdoor paths from A_t to S_{t+1} and from A_t to R_t.
- **Assumption 2:** For any $t < T$, $S_{t+1}, R_t \perp \{S_j, R_{j-1}, A_j\}_{j\le t-1} \mid S_t, A_t, Z$.
- **Assumption 3.** For any $t > 1$, U_t^S and U_{t-1}^R are deterministic functions of H_t.
- **Assumption 4 (additivity of exogeneous variables).** For all time $t \ge 0$, the exogeneous variables U_S^t and U_R^t are additive to S_t and R_t, respectively.

Algorithm 1 Proposed sequential data preprocessing

Input: Original data $\mathcal{D} = \{(s_{it}, z_i, a_{it}, r_{it}) : i = 1, \dots, N; t = 1, \dots, T\}$.

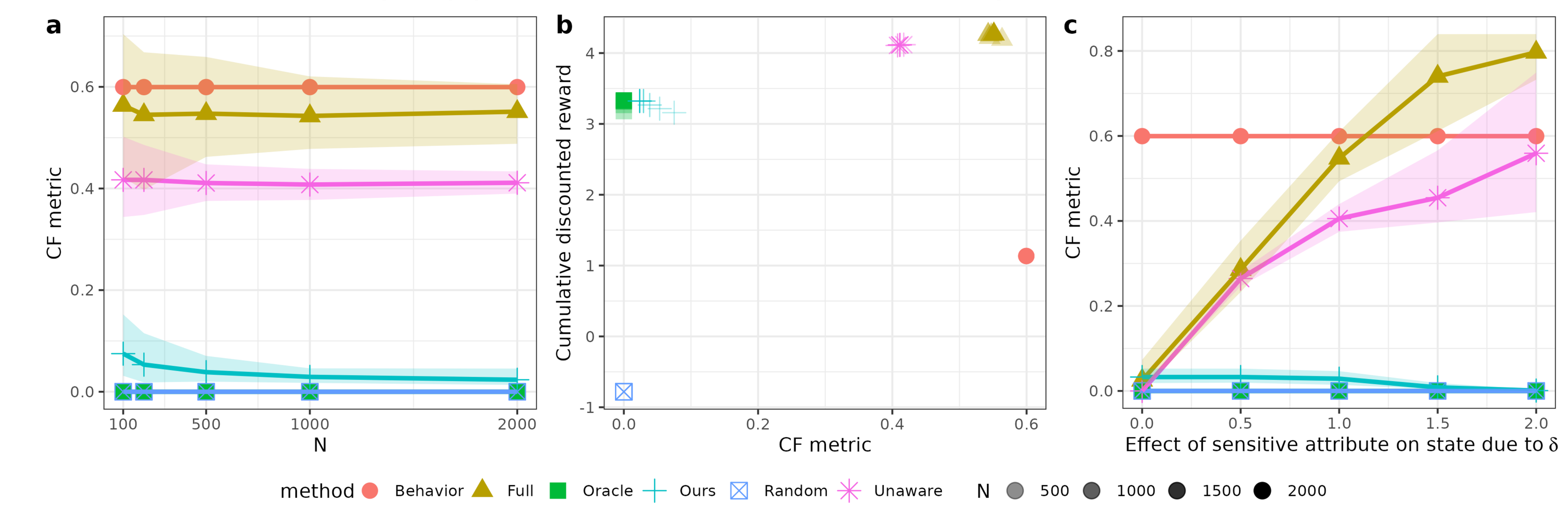
1: Fit the mean of transition kernel $\hat{\mu}(s, a, z)$ by MSE on $\mathcal{D}$.
2: Estimate $\hat{\mathbb{E}}(S_1|Z = z)$ and $\hat{P}(Z = z)$ $\forall z' \in \mathcal{Z}$ by the empirical means.
3: **for** $i = 1, \dots, N$ **do**
4: Calculate $\hat{s}_{i1}^{z'} = s_{i1} - \hat{\mathbb{E}}(S_1|Z = z) + \hat{\mathbb{E}}(S_1|Z = z'), \forall z' \in \mathcal{Z}$.
5: Set $\tilde{s}_{i1} = [\hat{s}_{i1}^{z^{(1)}}, \dots, \hat{s}_{i1}^{z^{(K)}}]^\top$.
6: **for** $t = 2, \dots, T$ **do**
7: $[\hat{s}_{it}^{z'}, \hat{r}_{i,t-1}^{z'}]^\top = [s_{it}, r_{i,t-1}]^\top - \hat{\mu}(s_{i,t-1}, a_{i,t-1}, z_i) + \hat{\mu}(\hat{s}_{i,t-1}^{z'}, a_{i,t-1}, z'), \forall z' \in \mathcal{Z}$.
8: $\tilde{s}_{it} = [\hat{s}_{it}^{z^{(1)}}, \dots, \hat{s}_{it}^{z^{(K)}}]^\top$,
9: $\tilde{r}_{i,t-1} = \sum_{k=1}^{K} \hat{P}(Z = z^{(k)}) \hat{r}_{i,t-1}^{z^{(k)}}$.
10: **end for**
11: **end for**

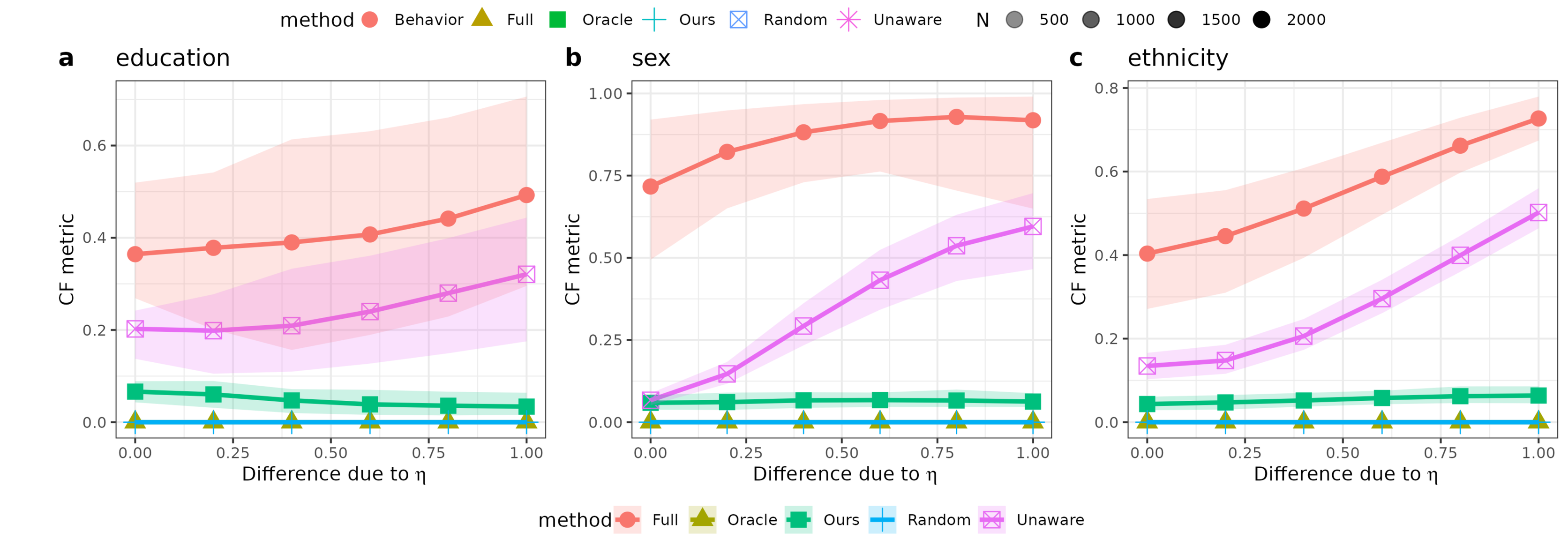
Output: Preprocessed experience tuples $\{(\tilde{s}_{it}, a_{it}, \tilde{r}_{it}) : i = 1, \dots, N; t = 1, \dots, T\}$.

calculate counterfactual states and rewards

Numerical Study

- Compare our proposal against the following in terms of value and fairness:
  - *Full:* uses all variables including the sensitive attribute - (S_t, Z).
  - *Unaware:* uses all variables except the sensitive attribute - (S_t).
  - *Oracle:* uses concatenations of counterfactual states and rewards, which are assumed to be known – $(\mathcal{S}_t)$.
  - *Random:* a policy that selects actions at random.
  - *Behavior:* the policy that was used to collect the input training data.
- We also investigate the impact of
  - Number of samples (N)
  - The strength (η) of the sensitive attribute's impact on states and rewards





Application to PowerED Study Data

- 207 patients over 12 weeks.
- sensitive attributes (separate analyses): education, age, sex, ethnicity.
- State variables: weekly pain, pain inference scores.
- $Reward = 7 - weekly\ self\ reported\ opioid\ medication\ risk\ score$
- Unfairness: Random < **Ours** < Unaware < Full
- Value: Full > Unaware > **Ours** > Random (in general)

Metric	Method	Education	Age	Sex	Ethnicity
Unfairness	Full	0.44 (0.14)	0.59 (0.15)	0.61 (0.15)	0.39 (0.13)
	Random	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Unaware	0.10 (0.03)	0.10 (0.02)	0.08 (0.03)	0.21 (0.05)
	Ours	**0.06 (0.02)**	**0.08 (0.02)**	**0.07 (0.02)**	**0.16 (0.03)**
Value	Full	57.09 (0.31)	57.29 (0.30)	57.20 (0.39)	56.87 (0.33)
	Random	56.61 (0.22)	56.66 (0.27)	56.53 (0.27)	56.54 (0.39)
	Unaware	57.01 (0.18)	57.21 (0.29)	56.96 (0.32)	57.00 (0.31)
	Ours	57.05 (0.30)	57.11 (0.28)	56.95 (0.51)	56.93 (0.48)

References

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems, 30.*

Chen, H., Lu, W., Song, R., & Ghosh, P. (2022). On Learning and Testing of Counterfactual Fairness through Data Preprocessing (No. arXiv:2202.12440). arXiv.

Wang, J., Shi, C., Piette, J.D., Loftus, J.R., Zeng, D. and Wu, Z., 2025. Counterfactually Fair Reinforcement Learning via Sequential Data Preprocessing. arXiv preprint arXiv:2501.06366.

With gratitude to Jianhan Zhang (jianhanz@umich.edu) for drafting this poster.