Achieving Long-term Fairness in Sequential Decision Making

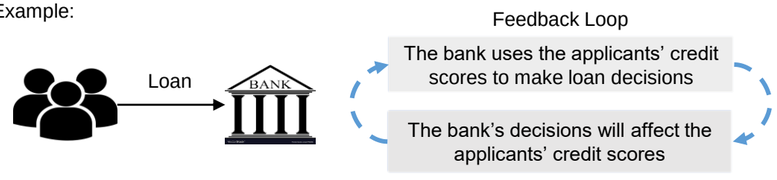
Yaowei Hu (yaoweihu@uark.edu) **Lu Zhang (**lz006@uark.edu)

University of Arkansas

Background

- Fair machine learning plays an important role in decision making tasks such as hiring, college admissions and bank loans.
- Fairness notions: *demographical parity*, *equalized odds* and *counterfactual fairness*.
- However, the majority of studies on fair machine learning focus on the static or one-shot classification setting.
- In practice, decision making systems are usually operating in a dynamic manner such as that the classifier makes sequential decision over a period of time.

Example:



Goal and Challenges

Our goal: Fair decision making should concern not only the fairness of a single decision but more importantly, whether a decision model can impose fair long-term effects on different groups. This notion of fairness is referred to as **long-term fairness**.

The **challenges** of achieving long-term fairness:

- **Feedback Loop**. Without knowing how the population would be reshaped by decisions, enforcing any fairness constraint may create negative feedback loops and eventually harm fairness in the long run.

- **Distribution Shift**. Ignoring the distribution shift will critically affect the achievement of long-term fairness, as long-term fairness is affected by all decisions made by the model along the time.

Causality-based Long-term Fairness

- **Definition 1** (**Long-term Fairness**). *The long-term fairness of a decision model* h_θ is measured by $P\left(\hat{Y}^{t*}(s_\pi^+,\theta)\right) - P\left(\hat{Y}^{t*}(s_\pi^-,\theta)\right)$ where π is a set of paths from S to $\hat{Y}^{t*}$ passing through $\boldsymbol{X}_r^1, \hat{Y}^1, \ldots, \boldsymbol{X}_r^{t*-1}, \hat{Y}^{t*-1}, \boldsymbol{X}_r^{t*}$, s_π represents the path-specific hard intervention and θ represents the soft intervention through all paths.
- **Definition 2** (**Short-term Fairness**). *The short-term fairness of a decision model* h_θ at time t is measured by the causal effect transmitted through paths involved in time t , i.e., $P\left(\hat{Y}^t(s_{\pi^t}^+,\theta)\right) - P\left(\hat{Y}^t(s_{\pi^t}^-,\theta)\right)$, where $\pi^t = \{S \to \widetilde{\boldsymbol{X}}_r \to \hat{Y}^t, S \to \hat{Y}^t\}$ with redlining attributes $\widetilde{\boldsymbol{X}}_r$, s_π is the path-specific hard intervention and θ represents the soft intervention.
- **Definition 3** (**Institute Utility**). *The institution utility of a decision model* h_θ is measured by the aggregate loss given by $\sum_{t=1}^{t*} E\left[L\left(Y^t, \hat{Y}^t\right)\right]$ where $L(\cdot)$ is the loss function.

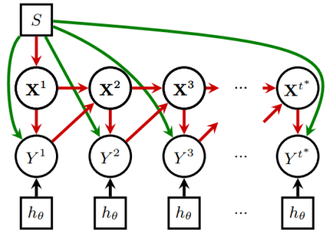
Problem Formulation 1. *The problem of fair sequential decision making is formulated as the constrained optimization:*

$$\arg\min_\theta \sum_{t=1}^{t*} E\left[L\left(Y^t, \hat{Y}^t\right)\right]$$

$$s.t.\ \ P\left(\hat{Y}^{t*}(s_\pi^+,\theta)=1\right) - P\left(\hat{Y}^{t*}(s_\pi^-,\theta)=1\right) \le \tau_l$$

$$P\left(\hat{Y}^t(s_{\pi^t}^+,\theta)=1\right) - P\left(\hat{Y}^t(s_{\pi^t}^-,\theta)=1\right) \le \tau_t,\ t=1,\ldots,t$$

where τ_l *and* τ_t *are thresholds for long-term and short term fairness constraints, respectively.*



Performative Risk Optimization

- To make it easier to solve the optimization problem in Problem Formulation 1, we convert it to the form of performative risk optimization.

Problem Formulation 2. *The problem of fair sequential decision making is reformulated as the performative risk optimization:*

$$\arg\min_\theta l(\theta) = \lambda_u l_u(\theta) + \lambda_l l_l(\theta) + \lambda_s l_s(\theta)$$

where λ_u, λ_l *and* λ_s *are weight parameters and satisfy* $\lambda_u + \lambda_l + \lambda_s = 1$.

$$l_u(\theta) = \sum_{t=1}^{t^*} \mathop{\mathbb{E}}_{S,\boldsymbol{X}^t,Y^t\sim P(S,\boldsymbol{X}^t,Y^t)} \left[\phi\left(Y^t h_\theta(\boldsymbol{X}^t, S)\right)\right],$$

$$l_l(\theta) = \frac{1}{2}\left\{ \mathop{\mathbb{E}}_{\boldsymbol{X}^{t^*}\sim P(\boldsymbol{X}^{t^*}\mid s_\pi^+,\theta)} \left[\phi\left(-h_\theta\left(\boldsymbol{X}^{t^*}, s^-\right)\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{X}^{t^*}\sim P(\boldsymbol{X}^{t^*}\mid s_\pi^-,\theta)} \left[\phi\left(h_\theta\left(\boldsymbol{X}^{t^*}, s^-\right)\right)\right] - 1 - \tau_l \right\},$$

$$l_s(\theta) = \frac{1}{t^*}\sum_{t=1}^{t^*} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{X}^t\sim P(\boldsymbol{X}^t\mid s_{\pi^t}^+,\theta)} \left[\phi\left(-h_\theta\left(\boldsymbol{X}^t, s^-\right)\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{X}^t\sim P(\boldsymbol{X}^t\mid s_{\pi^t}^-,\theta)} \left[\phi\left(h_\theta\left(\boldsymbol{X}^t, s^-\right)\right)\right] - 1 - \tau_t \right\}.$$

Repeated Risk Minimization

- **Repeated risk minimization** (RRM) is an iterative algorithmic heuristic for solving the performative risk optimization problem.

Theorem 1. *Suppose that surrogated loss function* $(\varphi \circ h)(\cdot)$ *is* β *–jointly smooth and* γ*-strongly convex and suppose that* $\boldsymbol{X}^{t+1}$ *are* c*-sensitive for any* t*, then the repeated risk minimization converges to a stable point at a linear rate, if* $2mc(t^*-1) < \frac{\beta}{\gamma}$.

Algorithm 1: Repeated Risk Minimization

Input : Dataset $\mathcal{D} = \{(S, \mathbf{X}^t, Y^t)\}_{t=1}^{t^*}$, time -lagged causal graph $\mathcal{G}$, convergence threshold δ
Output: The stable model h_θ

Train a classifier on $\mathcal{D}$ according to Eq. (2) without the soft intervention to obtain the initial parameter θ_0;
$i \leftarrow 0$;
repeat
  Sampled the post-intervention distributions $P\left(\mathbf{X}^{t^*}(s_\pi^+,\theta_i)\right)$ and $P\left(\mathbf{X}^{t^*}(s_\pi^-,\theta_i)\right)$;
  Sampled the post-intervention distributions $P\left(\mathbf{X}^t(s_{\pi^t}^+,\theta_i)\right)$ and $P\left(\mathbf{X}^t(s_{\pi^t}^-,\theta_i)\right)$ for each t;
  Minimize $l(\theta)$ according to Eq. (2) to obtain θ_{i+1};
  $\triangle \leftarrow \|\theta_{i+1} - \theta_i\|_2$;
  $i \rightarrow i+1$;
until $\triangle < \delta$;
$\theta \leftarrow \theta_i$;
return h_θ;

Experiments

Baselines:

- **Logistic Regression (LR)**: An unconstrained logistic regression model which takes user features and labels from all time steps as inputs and outputs.
- **Fair Model with Demographic Parity (FMDP)**: On the basis of the logistic regression model, fairness constraint is added to achieve demographic parity.
- **Fair Model with Equal Opportunity (FMEO):** On the basis of the logistic regression model, fairness constraint is added to achieve equal opportunity.

Data Generation Process:

We simulate a process of bank loans following the above time-lagged causal graph, where S is the protected attribute like race, $\boldsymbol{X}^t$ represents the financial status of applicants, and Y^t represents the decisions about whether to grant loans.

We sample the predicted decisions from:

$$P(\hat{Y}^t) = \sigma(h_{\theta^*}(\mathbf{X}^t, S)),\ \hat{Y}^t \sim 2 \cdot \text{Bernoulli}(P(\hat{Y}^t)) - 1.$$

$\boldsymbol{X}^{t+1}$ is generated according to the update rule below:

$$\mathbf{X}^{t+1} = \begin{cases} \mathbf{X}^t - \epsilon \cdot \theta^t + b & \hat{Y}^t = 1, Y^t = -1 \\ \mathbf{X}^t + \epsilon \cdot \theta^t + b & \hat{Y}^t = 1, Y^t = 1 \\ \mathbf{X}^t + b & \hat{Y}^t = -1 \end{cases}$$

Results of Synthetic and Semi-synthetic Datasets:

Table 1: Accuracy, short-term and long-term fairness of different algorithms on the synthetic dataset.

Alg.	Metric	Time steps				
		$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
RL	Acc	0.912	0.894	0.917	0.921	0.917
	Short	0.152	0.160	0.166	0.164	0.174
	Long	0.058	0.117	0.173	0.246	0.340
FMDP	Acc	0.735	0.706	0.704	0.708	0.725
	Short	0.212	0.216	0.224	0.220	0.232
	Long	0.180	0.306	0.376	0.431	0.481
FMEO	Acc	0.829	0.790	0.795	0.800	0.814
	Short	0.010	0.010	0.010	0.014	0.020
	Long	0.080	0.122	0.190	0.276	0.352
Ours	Acc	0.801	0.754	0.729	0.707	0.692
	Short	0.012	0.008	0.012	0.008	0.002
	Long	0.040	0.024	0.020	0.012	0.002

Table 2: Accuracy, short-term and long-term fairness of different algorithms on the semi-synthetic dataset.

Alg.	Metric	Time steps			
		$t=1$	$t=2$	$t=3$	$t=4$
RL	Acc	0.828	0.826	0.841	0.816
	Short	0.015	0.018	0.021	0.012
	Long	0.038	0.088	0.243	0.433
FMDP	Acc	0.830	0.843	0.846	0.841
	Short	0.063	0.066	0.075	0.069
	Long	0.038	0.076	0.223	0.397
FMEO	Acc	0.824	0.830	0.830	0.813
	Short	0.072	0.075	0.087	0.078
	Long	0.006	0.045	0.156	0.295
Ours	Acc	0.648	0.648	0.680	0.687
	Short	0.006	0.006	0.003	0.006
	Long	0.064	0.043	0.016	0.003

Acknowledgement

This work was supported in part by NSF 1910284, 1920920 and 1946391.