Equal Long-term Benefit Rate: Adapting Static Fairness Notions to Sequential Decision Making

Yuancheng Xu [*][1] **Chenghao Deng** [*][1] **Yanchao Sun** [1] **Ruijie Zheng** [1] **Xiyao Wang** [1] **Jieyu Zhao** [1] **Furong Huang** [1]

Abstract

Decisions made by machine learning models may have lasting impacts over time, making long-term fairness a crucial consideration. It has been shown that when ignoring the long-term effect of decisions, naively imposing fairness criterion in static settings can actually exacerbate bias over time. To explicitly address biases in sequential decision-making, recent works formulate long-term fairness notions in Markov Decision Process (MDP) framework. They define the long-term bias to be the sum of static bias over each time step. However, we demonstrate that naively summing up the step-wise bias can cause a false sense of fairness since it fails to consider the importance difference of states during transition. In this work, we introduce a new long-term fairness notion called ***Equal Long-term BEnefit RaTe*** (ELBERT), which explicitly considers state importance and can preserve the semantics of static fairness principles in the sequential setting. Moreover, we show that the policy gradient of Long-term Benefit Rate can be analytically reduced to standard policy gradient. This makes standard policy optimization methods applicable for reducing the bias, leading to our proposed bias mitigation method ELBERT-PO. Experiments on three dynamical environments show that ELBERT-PO successfully reduces bias and maintains high utility.

1. Introduction

The growing use of machine learning in decision making systems has raised concerns about potential biases to different sub-populations from underrepresented ethnicity, race, or gender (Dwork et al., 2012). In the real-world scenario, the decisions made by these systems can not only cause immediate unfairness, but can also have long-term effects on the future status of different groups. For example, in a loan application decision-making case, excessively denying loans to individuals from a disadvantaged group can have a negative impact on their future financial status and thus exacerbate the unfair inferior financial status in the long run.

It has been shown that when ignoring the long-term effects, naively imposing static fairness constraints such as demographic parity (Dwork et al., 2012) or equal opportunity (EO) (Hardt et al., 2016) can actually harm minorities (Liu et al., 2018; D'Amour et al., 2020). To explicitly address biases in sequential decision making problems, recent works (Wen et al., 2021; Chi et al., 2021; Yin et al., 2023) formulate the long-term effects in the framework of Markov Decision Process (MDP). MDP models the dynamics through the transition of states, e.g. how the number of applicants and their financial status change at the next time step given the current decisions. Also, MDP allows leveraging techniques in reinforcement learning (RL) for finding policies with better utility and fairness.

In sequential decision-making, states have different importance for fairness considerations. It is possible to transit from less important states to more important ones and vice versa. However, existing fairness criteria in the MDP framework fail to account for such difference. For example, consider the loan approval decision-making with two time steps and EO as the fairness criterion, as shown in Figure 1. For group blue, the state at time $\boldsymbol{t+1}$ is more important than time t, since there are more blue applicants at $\boldsymbol{t+1}$. For group red, state t is more important than $\boldsymbol{t+1}$. For group blue, the bank provides a high $\frac{\mathbf{100}}{\mathbf{100}}$ acceptance rate on a more important state $\boldsymbol{t+1}$ and a low $\frac{0}{1}$ acceptance rate on a less important state at t. However, for group red, the bank supplies a low $\frac{0}{100}$ acceptance rate on a more important state at t and a high $\frac{\mathbf{1}}{\mathbf{1}}$ acceptance rate on a less important state at $\boldsymbol{t+1}$. Therefore, group blue is more advantaged than group red, and bias emerges. In fact, overall, the bank makes an overall $\frac{100}{101}$ acceptance rate for group blue, much higher than $\frac{1}{101}$ for group red.

In a prior work (Yin et al., 2023), the authors define the

[*] Equal contribution [1] University of Maryland, College Park. Correspondence to: Yuancheng Xu <ycxu@umd.edu>.

2^{nd} *AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

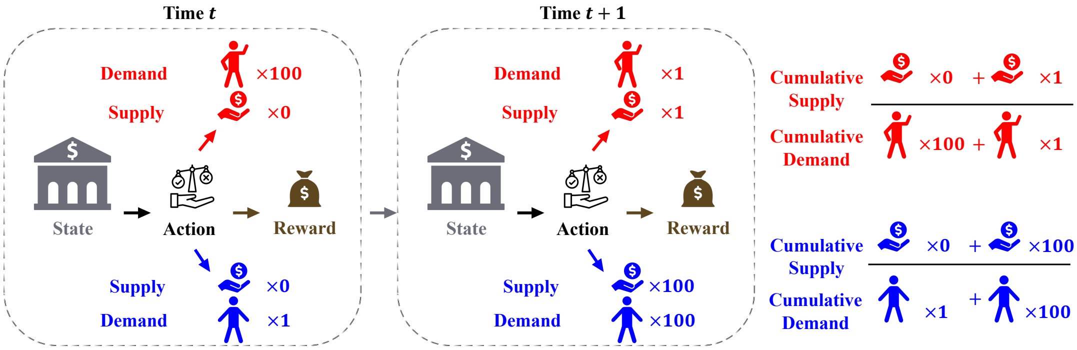


Figure 1. **(Left)** A loan application example with two groups in blue and red. At time step t, the bank approves 0 loans out of 1 qualified applicant from group blue and 0 loans out of 100 qualified applicants from group red. At time $\boldsymbol{t+1}$, the bank approves **100** loans out of **100** qualified applicants from group blue and **1** loans out of **1** qualified applicants from group red. **(Right)** The acceptance rate is 0 at time t and 1 at time $t+1$ for both groups, and thus the step-wise biases are zero and introduce a false sense of fairness. In contrast, our proposed Long-term Benefit Rate calculates the bias as $|\frac{1}{101} - \frac{100}{101}|$ and successfully identifies the bias.

long-term bias as the sum of step-wise bias (e.g. divergence of group acceptance rates), which, in the loan approval case, is calculated as $(\frac{0}{1} - \frac{0}{100})^2 + (\frac{\mathbf{100}}{\mathbf{100}} - \frac{\mathbf{1}}{\mathbf{1}})^2 = 0$. Another prior metric (Chi et al., 2021; Wen et al., 2021) defines the long-term bias as the difference of cumulative group rewards (e.g. acceptance rates) between two groups, i.e. $(\frac{0}{1} + \frac{\mathbf{100}}{\mathbf{100}}) - (\frac{0}{100} + \frac{\mathbf{1}}{\mathbf{1}}) = 0$. Neither of these metrics consider the state importance difference, i.e., the state with 100 qualified applicants is more important than the one with only 1 applicant. In fact, both metrics claim that there is no bias in this situation, leading to a false sense of fairness.

In this work, we introduce a new long-term fairness criterion called Equal Long-term Benefit Rate (ELBERT). Specifically, we define *Long-term Benefit Rate*, a general measure for the long-term well-being of a group, to be the ratio between the cumulative *group supply* (e.g. number of approved loans) and cumulative *group demand* (e.g. number of qualified applicants). For instance, in the loan application example, Long-term Benefit Rate calculates $\frac{100}{101}$ for group blue and $\frac{1}{101}$ for group red. By first summing up group supply and group demand separately and then taking the ratio, Long-term Benefit Rate takes into account that the group demand can vary over time steps. Thus ELBERT explicitly accounts for the change of state importance during transition, eliminating the false sense of fairness induced by prior metrics. Moreover, ELBERT is a general and versatile framework that can adapt several static fairness notions to their sequential setting counterparts through customization of group supply and group demand.

Furthermore, we propose a principled bias mitigation method, ELBERT Policy Optimization (ELBERT-PO), to reduce the differences of Long-term Benefit Rate among groups. Note that optimizing Long-term Benefit Rate is challenging since it is not in the standard form of cumulative reward in RL and how to compute its policy gradient was previously unclear. To address this, we show that the policy gradient of Long-term Benefit Rate can be analytically reduced to the standard policy gradient in RL. This makes efficient bias mitigation viable through adapting standard policy optimization methods. Experiments on three simulation environments show that our formulation and solution lead to significant improvement on group fairness while maintaining high utility.

Summary of Contributions. (1) We propose a new long-term fairness notion in the MDP setting, Equal Long-term Benefit Rate, which adapts static fairness notions and considers importance difference of states during transition. **(2)** We show that we can mitigate bias by manipulating Long-term Benefit Rate through adapting standard policy optimization methods. **(3)** Experimentally, we show that our bias mitigation method significantly improves fairness in sequential decision making.

2. ELBERT: Equal Long-term Benefit Rate for long-term fairness

Standard MDP Notations. A general sequential decision-making problem can be formulated as an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mu, T, R, \gamma \rangle$, where $\mathcal{S}$ is the state space (e.g. credit scores of applicants in the loan approval decision making mentioned above), μ is the initial state distribution, $\mathcal{A}$ is the action space (e.g. reject or approval), $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamic, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the immediate reward function (e.g. bank's earned profits) and γ is the discounting factor. The goal of RL is to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ to maximize cumulative reward $\eta(\pi) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$, where $s_0 \sim \mu$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim T(\cdot|s_t, a_t)$.

2.1. Supply-Demand Markov Decision Process for long-term fairness

Our goal is to formulate fairness in MDP, which requires defining the long-term well-being of each group. This motivates us to rethink the static notions of group well-being and how to adapt them to MDP.

Supply and demand in static settings. In many static fairness notions, the formulation of the group well-being can be unified as the ratio between supply and demand. For example, equal opportunity (EO)(Hardt et al., 2016) defines the well-being of group g as $\mathrm{P}[\hat{Y}=1|G=g,Y=1]=\frac{\mathrm{P}[\hat{Y}=1,Y=1,G=g]}{\mathrm{P}[Y=1,G=g]}$, where $\hat{Y}\in\{0,1\}$ is the binary decision (loan approval or rejection), $Y\in\{0,1\}$ is the target variable (repay or default) and G is the group ID. In practice, given a dataset, the well-being of group g, using the notion of EO, is calculated as $\frac{S_g}{D_g}$, where the supply S_g is the number of samples with $\{\hat{Y}=1,Y=1,G=g\}$ and the demand D_g is the number of samples with $\{Y=1,G=g\}$.

Note that such formulation in terms of supply and demand is not only restricted to EO, but is also compatible to other static fairness notions such as demographic parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016) and accuracy parity. We provide additional details in Appendix A.

Adapting to MDP. In the sequential setting, each time step corresponds to a static dataset that comes with group supply and group demand. Therefore, to adapt them to MDP, we assume that in addition to immediate reward $R(s_t,a_t)$, the agent receives immediate group supply $S_g(s_t,a_t)$ and immediate group demand $D_g(s_t,a_t)$ at every time step t. This is formalized as the Supply-Demand MDP (SD-MDP) as shown in Figure 2 and defined as follows.

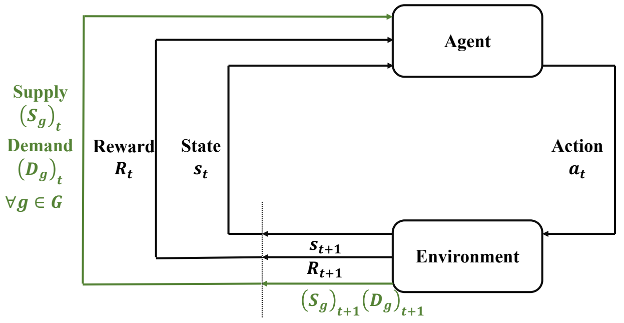


Figure 2. Supply Demand MDP (SD-MDP). In addition to the standard MDP (in black), SD-MDP returns group demand and group supply as fairness signals (in green).

Definition 2.1 (Supply-Demand MDP (SD-MDP))**.** Given a group index set G and a standard MDP $\mathcal{M}=\langle\mathcal{S},\mathcal{A},\mu,T,R,\gamma\rangle$, a Supply-Demand MDP is $\mathcal{M}_{\mathrm{SD}}=\langle\mathcal{S},\mathcal{A},\mu,T,R,\gamma,\{S_g\}_{g\in G},\{D_g\}_{g\in G}\rangle$. Here $\{S_g\}_{g\in G}$ and $\{D_g\}_{g\in G}$ are immediate group supply and group demand function for group g.

Compared with the standard MDP, in SD-MDP, an agent receives additional fairness signals $S_g(s_t,a_t)$ and $D_g(s_t,a_t)$ after taking action a_t at each time step. To characterize the long-term group supply and group demand of a policy π, we define cumulative group supply and group demand as follows.

Definition 2.2 (Cumulative Supply and Demand)**.** Define the cumulative group supply as $\eta_g^S(\pi):=\mathbb{E}_\pi\left[\sum_{t=0}^\infty\gamma^t S_g(s_t,a_t)\right]$ and cumulative group demand as $\eta_g^D(\pi):=\mathbb{E}_\pi\left[\sum_{t=0}^\infty\gamma^t D_g(s_t,a_t)\right]$.

2.2. Proposed long-term fairness metric: Equal Long-term Benefit Rate (ELBERT)

In the following definitions, we propose to measure the well-being of a group by the ratio of cumulative group supply and group demand and propose the corresponding fairness metric: Equal Long-term Benefit Rate (ELBERT).

Definition 2.3 (Long-term Benefit Rate)**.** Define the Long-term Benefit Rate of group g as $\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}$. Define the bias of a policy as the maximal difference of Long-term Benefit Rate among groups, i.e., $b(\pi)=\max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}-\min_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}$.

RL with ELBERT. Under the framework of ELBERT, the goal of reinforcement learning with fairness constraints is to find a policy to maximize the cumulative reward and keep the bias under a threshold ϵ. In other words,

$$\max_\pi \eta(\pi) \quad \text{s.t.} \quad b(\pi)=\max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}-\min_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}\le\epsilon. \tag{1}$$

Relationship with static fairness notions. Note that in the special case when the length of time horizon is 1, Long-term Benefit Rate reduces to $\frac{S_g}{D_g}$, i.e., the static fairness notion.

Versatility. By choosing the proper definition of group supply S_g and group demand D_g according to the static fairness notion, Equal Long-term Benefit Rate is customized to adapt the static notion to sequential decision-making.

Comparison to other fairness metrics in MDP. The fairness notion called return parity proposed in previous work (Wen et al., 2021; Chi et al., 2022) use cumulative individual rewards to measure the group well-being. It can be viewed as a special case of Long-term Benefit Rate with the demand function $D_g(s,a)$ being a constant function, and ignoring the importance difference of states during transition. As demonstrated in Section 1, this metric can cause a false sense of fairness.

3. Achieving Equal Long-term Benefit Rate

In this section, we will develop a bias mitigation algorithm, ELBERT Policy Optimization (ELBERT-PO) to solve the

RL problem with the fairness considerations in Equation (1). In Section 3.1, we will formulate the training objective as a policy optimization problem and lay out the challenge of computing the policy gradient of this objective. In Section 3.2, we demonstrate how to compute the policy gradient of this objective by reducing it to standard policy gradient. In Section 3.3, we extend the objective and its solution to multi-group setting and deal with the non-smoothness of the maximum and minimum operator.

3.1. Training objective and its challenge

Objective. We first consider the special case of two groups $G = \{1, 2\}$, where Long-term Benefit Rate reduces to $|\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)} - \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)}|$. To solve the constrained problem in Equation (1), we propose to solve the unconstrained relaxation of it by maximizing the following objective:

$$J(\pi) = \eta(\pi) - \alpha b(\pi)^2 = \eta(\pi) - \alpha(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)} - \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)})^2 \quad (2)$$

where α is a constant controlling the trade-off between the total return and the bias.

Challenge: policy gradient of $b(\pi)$. To optimize the objective above, it is natural to use policy optimization methods that estimate the policy gradient and use stochastic gradient ascent to directly improve policy performance. However, in order to compute the policy gradient $\nabla_\pi J(\pi)$ of the objective function $J(\pi)$ in Equation (2), one needs to compute $\nabla_\pi \eta(\pi)$ and $\nabla_\pi b(\pi)$. Although the term $\nabla_\pi \eta(\pi)$ is a standard policy gradient that has been extensively studied in RL(Schulman et al., 2016), it was previously unclear how to deal with $\nabla_\pi b(\pi) = \nabla_\pi(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)} - \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)})$. In particular, since $b(\pi)$ is not of the form of expected total return, one cannot directly apply Bellman Equation to compute $b(\pi)$. Therefore, it is unclear how to leverage standard policy optimization methods(Schulman et al., 2017; 2015) to the objective function $J(\pi)$.

3.2. Solution to the objective

In this section, we show how to apply existing policy optimization methods to solve the objective in Equation (2). This is done by analytically reducing the policy gradient $\nabla_\pi b(\pi)$ of the bias to standard policy gradients.

Gradient of the objective. For the simplicity of notation, we denote the term $b(\pi)^2$ in Equation (2) as a function of Long-term Benefit Rate $\{\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}\}_{g\in G}$ as $b(\pi)^2 = h(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)}, \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)})$, where $h(z_1, z_2) = (z_1 - z_2)^2$. Therefore, $J(\pi) = \eta(\pi) - h(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)}, \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)})$. By chain rule, we can compute the gradient of the objective as follows.

$$\nabla_\pi J(\pi) = \nabla_\pi \eta(\pi) - \alpha \sum_{g\in G} \frac{\partial h}{\partial z_g} \nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) \quad (3)$$

where $\frac{\partial h}{\partial z_g}$ is the partial derivative of h w.r.t. its g-th coordinate, evaluated at $(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)}, \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)})$. Note that $\nabla_\pi \eta(\pi)$ in Equation (3) is a standard policy gradient, whereas $\nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$ is not.

Reduction to standard policy gradient. To estimate $\nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$, we apply the chain rule again as follows

$$\nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) = \frac{1}{\eta_g^D(\pi)} \nabla_\pi \eta_g^S(\pi) - \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)^2} \nabla_\pi \eta_g^D(\pi) \quad (4)$$

Therefore, in order to estimate $\nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$, one only needs to estimate the expected total supply and demand $\eta_g^S(\pi), \eta_g^D(\pi)$ as well as the standard policy gradients $\nabla_\pi \eta_g^S(\pi), \nabla_\pi \eta_g^D(\pi)$.

Advantage function for policy gradient. It is common to compute a policy gradient $\nabla_\pi \eta(\pi)$ using $\mathbb{E}_\pi\{\nabla_\pi \log \pi(a_t|s_t) A_t\}$, where A_t is the advantage function of the reward R. Denote the advantage functions of $R, \{S_g\}_{g\in G}, \{D_g\}_{g\in G}$ as $A_t, \{A_{g,t}^S\}_{g\in G}, \{A_{g,t}^D\}_{g\in G}$. $\nabla_\pi(\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$ in Equation (4) can thus be written as

$$\mathbb{E}_\pi\left\{\nabla_\pi \log \pi(a_t|s_t)(\frac{1}{\eta_g^D(\pi)} A_{g,t}^S - \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)^2} A_{g,t}^D)\right\} \quad (5)$$

By plugging Equation (5) into Equation (3), we obtain the gradient of the objective $J(\pi)$ using advantage functions as follows

$$\nabla_\pi J(\pi) = \mathbb{E}_\pi\{\nabla_\pi \log \pi(a_t|s_t) A_t^{\text{fair}}\} \quad (6)$$

Therefore, $\nabla_\pi J(\pi) = \mathbb{E}_\pi\{\nabla_\pi \log \pi(a_t|s_t) A_t^{\text{fair}}\}$, where $A_t^{\text{fair}} = A_t - \alpha \sum_{g\in G} \frac{\partial h}{\partial z_g}(\frac{1}{\eta_g^D(\pi)} A_{g,t}^S - \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)^2} A_{g,t}^D)$ is defined as the *fairness-aware advantage* function. In practice, we use PPO (Schulman et al., 2017) with the fairness-aware advantage function A_t^{fair} to update the policy network for better training stability. The resulting algorithm ELBERT Policy Optimization (ELBERT-PO), is given in Algorithm 1. In particular, in line 11-13, PPO objective $J^{\text{CLIP}}(\theta)$ is used, where $\hat{\mathbb{E}}_{\pi_\theta}$ denotes the empirical average over samples collected by π_θ and ϵ is a hyperparameter for clipping.

3.3. Extension to multi-group setting

Challenge: Non-smoothness in multi-group bias. When there are multiple groups, the objective is $J(\pi) = \eta(\pi) -$

Algorithm 1 ELBERT Policy Optimization (ELBERT-PO)

```
Input: Group set $G$, bias trade-off factor $\alpha$, bias function $h$, temperature $\beta$ (if multi-group)
Initialize policy network $\pi_\theta(a|s)$, value networks $V_\phi(s)$, $V_{\phi_g^S}(s)$, $V_{\phi_g^D}(s)$ for all $g \in G$
for $k \leftarrow 0, 1, ...$ do
    Collect a set of trajectories $\mathcal{D} \leftarrow \{\tau_k\}$ by running $\pi_\theta$ in the environment, each trajectory $\tau_k$ contains $\tau_k :\leftarrow \{(s_t, a_t, r_t, s_{t+1})\}, t \in [|\tau_k|]$
    Compute the cumulative rewards, supply and demand $\eta, \eta_g^S, \eta_g^D$ of $\pi_\theta$ using Monte Carlo
    for each gradient step do
        Sample a mini-batch from $\mathcal{D}$
        Compute advantages $A_t, A_{g,t}^S, A_{g,t}^D$ using the current value networks $V_\phi(s)$, $V_{\phi_g^S}(s)$, $V_{\phi_g^D}(s)$ and mini-batch for all $g \in G$
        Compute $\frac{\partial h}{\partial z_g}$ at $(\frac{\eta_1^S}{\eta_1^D}, \cdots, \frac{\eta_M^S}{\eta_M^D})$
        Compute the fairness-aware advantage function:
            $$A_t^{\text{fair}} = A_t - \alpha \sum_{g \in G} \frac{\partial h}{\partial z_g} (\frac{1}{\eta_g^D} A_{g,t}^S - \frac{\eta_g^S}{(\eta_g^D)^2} A_{g,t}^D)$$
        $R_t(\theta) \leftarrow \pi_\theta(s_t, a_t) / \pi_{\theta_{\text{old}}}(s_t, a_t)$
        $J^{\text{CLIP}}(\theta) \leftarrow \hat{\mathbb{E}}_{\pi_\theta}[\min(R_t(\theta) A_t^{\text{fair}}, \text{clip}(R_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t^{\text{fair}})]$
        Update the policy network $\theta \leftarrow \theta + \tau \nabla_\theta J^{\text{CLIP}}(\theta)$
        Fit $V_\phi(s)$, $V_{\phi_g^S}(s)$, $V_{\phi_g^D}(s)$ by regression on the mean-squared error
    end for
end for
```

$\alpha b(\pi)^2 = \eta(\pi) - \alpha(\max_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} - \min_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})^2$. However, the max and min operator can cause non-smoothness in the objective during training. This is because only the groups with the maximal and minimal Long-term Benefit Rate will affect the bias term and thus the gradient of it. This is problematic especially when there are several other groups with Long-term Benefit Rate close to the maximal or minimal values. The training algorithm should consider all groups and decrease all the high Long-term Benefit Rate and increase low ones.

Soft bias in multi-group setting. To solve this, we replace the max and min operator in $b(\pi)$ with their smoothed version controlled by the temperature $\beta > 0$ and define the soft bias $b^{\text{soft}}(\pi)$ as $\frac{1}{\beta} \log \sum_{g \in G} \exp(\beta \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) - \frac{1}{-\beta} \log \sum_{g \in G} \exp(-\beta \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$. The relationship between the exact and soft bias is characterized by the following:

Proposition 3.1. *Given a policy π, the number of groups M and the temperature β, $b(\pi) \leq b^{\text{soft}}(\pi) \leq b(\pi) + \frac{2 \log M}{\beta}$.*

In other words, the soft bias is an upper bound of the exact bias and moreover, the quality of such approximation is controllable: the gap between the two decreases as β increases and vanishes when $\beta \to \infty$. We provide the proof in Appendix B. Therefore, we maximize $J(\pi) = \eta(\pi) - \alpha b^{\text{soft}}(\pi)^2$ in the multi-group settings. We write $b^{\text{soft}}(\pi)^2 = h(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)}, \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)}, ..., \frac{\eta_M^S(\pi)}{\eta_M^D(\pi)})$ where $h(z) = [\frac{1}{\beta} \log \sum_g \exp(\beta z_g) - \frac{1}{-\beta} \log \sum_g \exp(-\beta z_g)]^2$, $z = (z_1, \cdots, z_M)$ and $J(\pi) = \eta(\pi) - h(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)}, \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)}, ..., \frac{\eta_M^S(\pi)}{\eta_M^D(\pi)})$. The gradient $\nabla_\pi J(\pi)$ is still computed by Equation (3) and the training pipeline still follows Algorithm 1.

4. Related Work

Fairness criterion in MDP. A line of work has formulated fairness in the framework of MDP. The work in (D'Amour et al., 2020) proposes to study long-term fairness in MDP using simulation environments and shows that static fairness notions can contradict with long-term fairness. Return parity (Chi et al., 2022; Wen et al., 2021) assumes that the long-term group benefit can be represented by the sum of group benefit at each time step. However, as illustrated in Section 1, this assumption is problematic since it ignores the importance difference among states during transition. The work (Yin et al., 2023) formulates the long-term bias as the sum of static bias at each time steps, suffering from the same problem. Our proposed ELBERT explicitly considers the importance difference among states through the SD-MDP. Another work (Yu et al., 2022) assumes that there exists a long-term fairness measure for each state and proposes A-PPO, a advantage regularized policy optimization method to encourage the bias at the next time step to be smaller than the bias at the current time step. However, the assumption of (Yu et al., 2022) does not hold in general since for a trajectory that ends with a certain state, the long-term fairness depends on the whole history of state-action pairs instead of only a single state. Moreover, APPO only encourages the bias of the next time step to be smaller than the current one, whereas our proposed ELBERT-PO considers the bias in all future steps.

Long-term fairness in other temporal models. Long-term fairness is also studied in other temporal models. The work in (Liu et al., 2018) shows that naively imposing static fairness constraints in a one-step feedback model can actually harm the minority, showing the necessity of explicitly accounting for sequential decisions. Effort-based fairness (Heidari et al., 2019; Guldogan et al., 2022) measures bias as the disparity in the effort made by individuals from each group to get a target outcome, where the effort only considers one future time step. Long-term fairness has also been studied in multi-armed bandit (Chen et al., 2020), which do not consider how decisions influences the state of the environment. In this work, we study long-term fairness in MDP since it is a general framework to model the dynamics in real world and allows leveraging existing RL techniques for finding high-utility policy with fairness constraints.

5. Experiment

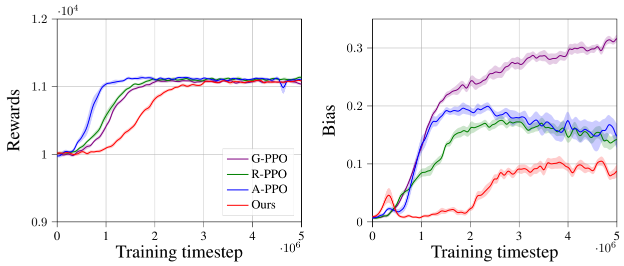
In this section, we investigate the effectiveness of the proposed ELBERT-PO in the loan approval environment (D'Amour et al., 2020). In Appendix C.3, we also demonstrate the performance of ELBERT-PO in a multi-group setting in the attention allocation environment.

Lending environment. In this environment, a bank (the RL agent) decides whether to accept or reject loan applications and the applicants arrive one at a time sequentially. There are two groups among applicants ($G = \{1, 2\}$). The applicant at each time t is from one of the groups g_t and has a credit score sampled from the credit score distribution of group g_t. A higher credit score means higher repaying probability if the loan is approved. Group 2 is disadvantaged with a lower mean of the initial credit score distribution compared with Group 1. As for the dynamics, at time t, the credit score distribution of group g_t shifts higher if its group member gets loan approval (i.e. $\hat{Y}_t = 1$) and repays the loan (i.e. $Y_t = 1$). The immediate reward is the increment of the bank cash at each time step.

Bias metric. The bias is defined by $\left|\frac{\sum_t \mathbb{1}\{G_t=0,Y_t=\hat{Y}_t=1\}}{\sum_t \mathbb{1}\{G_t=0,Y_t=1\}} - \frac{\sum_t \mathbb{1}\{G_t=1,Y_t=\hat{Y}_t=1\}}{\sum_t \mathbb{1}\{G_t=1,Y_t=1\}}\right|$, which is the long-term extension of EO, where the group well-being is measured by the true positive rate. In the Long-term Benefit Rate framework, group supply $D_g(s_t, a_t) = \mathbb{1}\{G_t = g, Y_t = \hat{Y}_t = 1\}$ and group demand $S_g(s_t, a_t) = \mathbb{1}\{G_t = g, Y_t = 1\}$.

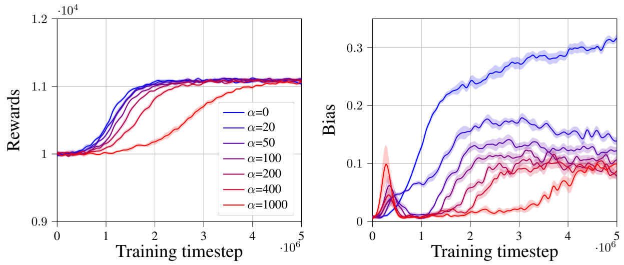
Baselines. Following Yu et al. (2022), we consider the following RL baselines. (1) A-PPO (Yu et al., 2022), which regularizes the advantage function to decrease the bias of the next time steps but does not consider the biases in all future steps. (2) Greedy PPO (G-PPO), which greedily maximizes reward without any fairness considerations. (3) Reward-Only Fairness Constrained PPO (R-PPO), a heuristic method which directly injects the bias of all previous time steps into the current reward. We list all the hyperparameters settings in Appendix C.2.

Results: ELBERT-PO reduces the bias while maintaining high reward. We present the learning curve of ELBERT-PO and baselines in Figure 3. In the lending experiment, ELBERT-PO reduces the bias to 0.09, which significantly decreases the bias of G-PPO by about 70% and is about 33% lower than A-PPO and R-PPO. This demonstrates the importance of formulating the long-term fairness criterion using cumulative demand and supply as well as considering the biases of all future time steps during training, as done by ELBERT-PO. Also, ELBERT-PO converges to obtain the same reward as baselines. Therefore, it shows that in the loan application environment, ELBERT-PO successfully reduces the bias while still attaining high reward.



(a) Rewards (b) Bias

Figure 3. Learning curve for the loan application environment.



(a) Rewards (b) Bias

Figure 4. Learning curve of ELBERT-PO with different α.

Effect of α. In Figure 4, the learning curve with different values of α is shown. We observe that larger α leads to lower bias, though such effect is diminishing as α becomes larger. In terms of reward, we find that increasing α leads to slower convergence. This is expected since the reward signal becomes weaker as α increases. However, we find that ELBERT-PO on all considered α values converge to the same reward value. This suggests that lower bias does not necessarily leads to lower rewards.

6. Conclusions and discussions

In this work, we introduce a new long-term fairness notion called Equal Long-term Benefit Rate (ELBERT). ELBERT explicitly accounts for the varying state importance in sequential decision-making through the Supply-Demand MDP. We analytically reduce the policy gradient of Long-term Benefit Rate to standard policy gradient, which leads to the ELBERT-PO method for bias mitigation. Experimental results demonstrate that ELBERT-PO successfully reduces bias while maintaining high utility. One limitation is that ELBERT-PO uses on-policy RL methods and might suffer from poor sample complexity. Future work includes designing off-policy algorithms for ELBERT-PO to improve the sample complexity.

References

Atwood, J., Srinivasan, H., Halpern, Y., and Sculley, D. Fair treatment allocations in social networks. *arXiv preprint arXiv:1911.05489*, 2019.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., and Nikolaidis, S. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020.

Chi, J., Tian, Y., Gordon, G. J., and Zhao, H. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning*, pp. 1866–1876. PMLR, 2021.

Chi, J., Shen, J., Dai, X., Zhang, W., Tian, Y., and Zhao, H. Towards return parity in markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1161–1178. PMLR, 2022.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M. J., Neel, S., Roth, A., and Schutzman, Z. Fair algorithms for learning in allocation problems. In danah boyd and Morgenstern, J. H. (eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 170–179. ACM, 2019. doi: 10.1145/3287560.3287571. URL https://doi.org/10.1145/3287560.3287571.

Guldogan, O., Zeng, Y., Sohn, J.-y., Pedarsani, R., and Lee, K. Equal improvability: A new fairness notion considering the long-term impact. *arXiv preprint arXiv:2210.06732*, 2022.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Heidari, H., Nanda, V., and Gummadi, K. P. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Wen, M., Bastani, O., and Topcu, U. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021.

Xu, Y., Sun, Y., Goldblum, M., Goldstein, T., and Huang, F. Exploring and exploiting decision boundary dynamics for adversarial robustness. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.03015.

Yin, T., Raab, R., Liu, M., and Liu, Y. Long-term fairness with unknown dynamics. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. URL https://openreview.net/forum?id=ZswoPbTlNQ.

Yu, E. Y., Qin, Z., Lee, M. K., and Gao, S. Policy optimization with advantage regularization for long-term fairness in decision systems. *arXiv preprint arXiv:2210.12546*, 2022.

Supplementary Material

A. Fairness notions with the supply and demand formulation

In this section, we demonstrate that in the static settings, the supply and demand formulation in Section 2 can cover many popular fairness notions. This means that the proposed Supply Demand MDP is expressive enough to extend several popular static fairness notions to the sequential settings. In the following, we give a list of examples to show, in the static setting, how to formulate several popular fairness criteria as the ratio between the supply and demand. For simplicity, we consider the agent's decision to be binary, though the analysis naturally extends to multi-class settings.

Notations. Denote $\hat{Y} \in \{0, 1\}$ as the binary decision (loan approval or rejection), $Y \in \{0, 1\}$ as the target variable (repay or default) and G as the group ID.

Demographic Parity. The well-being of a group g in Demographic Parity (DP) (Dwork et al., 2012) is defined as $\mathrm{P}[\hat{Y} = 1|G = g] = \frac{\mathrm{P}[\hat{Y}=1,G=g]}{\mathrm{P}[G=g]}$ and DP requires such group well-being to equalized among groups. In practice, given a dataset, the well-being of group g is calculated as $\frac{S_g}{D_g}$, where the supply S_g is the number of samples with $\{\hat{Y} = 1, G = g\}$ (e.g. the number of accepted individuals in group g) and the demand D_g is the number of samples with $\{G = g\}$ (e.g. the total number of individuals from group g).

Equal Opportunity. The well-being of a group g in Equal Opportunity (EO) (Dwork et al., 2012) is defined as $\mathrm{P}[\hat{Y} = 1|G = g, Y = 1] = \frac{\mathrm{P}[\hat{Y}=1,Y=1,G=g]}{\mathrm{P}[Y=1,G=g]}$ and EO requires such group well-being to equalized among groups. In practice, given a dataset, the well-being of group g is calculated as $\frac{S_g}{D_g}$, where the supply S_g is the number of samples with $\{\hat{Y} = 1, Y = 1, G = g\}$ (e.g. the number of qualified and accepted individuals in group g) and the demand D_g is the number of samples with $\{Y = 1, G = g\}$ (e.g. the number of qualified individuals from group g).

Equality of discovery probability: a special case of EO Equality of discovery probability (Elzayn et al., 2019) requires that the discovery probability to be equal among groups. For example, in predictive policing setting, it requires that conditional on committing a crime ($Y = 1$), the probability that an individual is apprehended ($\hat{Y} = 1$) should be independent of the district ID (group ID) g. This is a special case of EO in specific application settings.

Equalized Odds. Equalized Odds (Dwork et al., 2012) requires that both the True Positive Rate (TPR) $\mathrm{P}[\hat{Y} = 1|G = g, Y = 1] = \frac{\mathrm{P}[\hat{Y}=1,Y=1,G=g]}{\mathrm{P}[Y=1,G=g]}$ and the False Positive Rate (FPR) $\mathrm{P}[\hat{Y} = 1|G = g, Y = 0] = \frac{\mathrm{P}[\hat{Y}=1,Y=0,G=g]}{\mathrm{P}[Y=0,G=g]}$ equalize among groups. In practice, given a dataset, **(a)** the TPR of group g is calculated as $\frac{S_g^T}{D_g^T}$, where the supply S_g^T is the number of samples with $\{\hat{Y} = 1, Y = 1, G = g\}$ (e.g. the number of qualified and accepted individuals in group g) and the demand D_g^T is the number of samples with $\{Y = 1, G = g\}$ (e.g. the number of qualified individuals from group g). **(b)** The FPR of group g is calculated as $\frac{S_g^F}{D_g^F}$, where the supply S_g^F is the number of samples with $\{\hat{Y} = 1, Y = 0, G = g\}$ (e.g. the number of unqualified but accepted individuals in group g) and the demand D_g^F is the number of samples with $\{Y = 0, G = g\}$ (e.g. the number of unqualified individuals from group g).

Extending Equalized Odds to sequential settings using SD-MDP. The long-term adaption of Equalized Odds can be included by the Supply Demand MDP via allowing it to have two sets of supply-demand pairs: for every group g, (D_g^T, S_g^T) and (D_g^F, S_g^F). In particular, define the cumulative supply and demand for both supply-demand pairs: the cumulative group supply for TPR $\eta_g^{S,T}(\pi) := \mathbb{E}_\pi \big[\sum_{t=0}^{\infty} \gamma^t S_g^T(s_t, a_t) \big]$ and cumulative group demand for TPR as $\eta_g^{D,T}(\pi) := \mathbb{E}_\pi \big[\sum_{t=0}^{\infty} \gamma^t D_g^T(s_t, a_t) \big]$. The cumulative group supply for FPR $\eta_g^{S,F}(\pi) := \mathbb{E}_\pi \big[\sum_{t=0}^{\infty} \gamma^t S_g^F(s_t, a_t) \big]$ and cumulative group demand for FPR as $\eta_g^{D,F}(\pi) := \mathbb{E}_\pi \big[\sum_{t=0}^{\infty} \gamma^t D_g^F(s_t, a_t) \big]$. Since the bias considers both TPR and FPR, we define the bias for both: $b^T(\pi) = \max_{g \in G} \frac{\eta_g^{S,T}(\pi)}{\eta_g^{D,T}(\pi)} - \min_{g \in G} \frac{\eta_g^{S,T}(\pi)}{\eta_g^{D,T}(\pi)}$ and $b^F(\pi) = \max_{g \in G} \frac{\eta_g^{S,F}(\pi)}{\eta_g^{D,F}(\pi)} - \min_{g \in G} \frac{\eta_g^{S,F}(\pi)}{\eta_g^{D,F}(\pi)}$. The goal of RL with Equalized Odds constraints can be formulated as

$$\begin{aligned}
&\max_{\pi} \eta(\pi) \\
&\text{s.t. } b^T(\pi) = \max_{g\in G} \frac{\eta_g^{S,T}(\pi)}{\eta_g^{D,T}(\pi)} - \min_{g\in G} \frac{\eta_g^{S,T}(\pi)}{\eta_g^{D,T}(\pi)} \le \epsilon \\
&\quad b^F(\pi) = \max_{g\in G} \frac{\eta_g^{S,F}(\pi)}{\eta_g^{D,F}(\pi)} - \min_{g\in G} \frac{\eta_g^{S,F}(\pi)}{\eta_g^{D,F}(\pi)} \le \epsilon.
\end{aligned} \tag{7}$$

In practice, we treat the hard constraints as regularization and use the following objective function

$$J(\pi) = \eta(\pi) - \alpha b^T(\pi)^2 - \alpha b^F(\pi)^2 \tag{8}$$

where α is a trade-off constant between return and fairness. The gradient $\nabla_\pi(J\pi)$ can still be computed using techniques presented in 3, since both bias terms $b^T(\pi)$ and $b^F(\pi)$ are still in the form of ratio between cumulative supply and demand.

Accuracy Parity Accuracy Parity defines the well-being of group g as $\mathrm{P}[\hat{Y} = Y|G = g] = \frac{\mathrm{P}[\hat{Y}=Y,G=g]}{\mathrm{P}[G=g]}$, which is the accuracy of predicting Y using $\hat{Y}$ among individuals from the group g. In practice, this is computed by $\frac{S_g}{D_g}$, where the supply S_g is the number of samples with $\{\hat{Y} = Y, G = g\}$ (e.g. the number of individuals with correct predictions in group g) and the demand D_g is the number of samples with $\{G = g\}$ (e.g. the total number of individuals from group g).

B. Relationship between the soft bias and the bias

We would like to show the mathematical relationship between the soft bias and bias, as shown in Proposition 3.1. This is done by analyzing the max and min operator as well as their soft counterparts through the log sum trick, which is also explored in prior work (Xu et al., 2023). We restate the full proposition and present the proof below.

Proposition B.1. *Given a policy π, the number of groups M and the temperature β, define the soft bias as*

$$b^{soft}(\pi) = \frac{1}{\beta}\log\sum_{g\in G}\exp(\beta\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) - \frac{1}{-\beta}\log\sum_{g\in G}\exp(-\beta\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}).$$

The bias is defined as

$$b(\pi) = \max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} - \min_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}.$$

We have that

$$b(\pi) \le b^{soft}(\pi) \le b(\pi) + \frac{2\log M}{\beta}.$$

Proof. First consider the first term $\frac{1}{\beta}\log\sum_{g\in G}\exp(\beta\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)})$ in the soft bias $b^{\text{soft}}(\pi)$.

On the one hand, we have that

$$\begin{aligned}
\frac{1}{\beta}\log\sum_{g\in G}\exp(\beta\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) &> \frac{1}{\beta}\log\exp(\beta\max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) \\
&= \max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}
\end{aligned} \tag{9}$$

On the other hand, we have that

$$\begin{aligned}
\frac{1}{\beta}\log\sum_{g\in G}\exp(\beta\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) &\le \frac{1}{\beta}\log M\exp(\beta\max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) \\
&= \max_{g\in G}\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} + \frac{\log M}{\beta}
\end{aligned} \tag{10}$$

Therefore, $\max_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} < \frac{1}{\beta} \log \sum_{g \in G} \exp(\beta \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) \leq \max_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} + \frac{\log M}{\beta}$.

Similarly, it can be shown that $\min_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} - \frac{\log M}{\beta} \leq \frac{1}{-\beta} \log \sum_{g \in G} \exp(-\beta \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}) < \min_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}$.

By subtracting the two, we conclude that $b(\pi) \leq b^{\text{soft}}(\pi) \leq b(\pi) + \frac{2 \log M}{\beta}$.

□

C. Experimental Details

C.1. Full description of the environments

Lending In the first environment, we consider credit approval for lending in a sequential setting (Liu et al., 2018). As the agent in this scenario, a bank decides whether to accept or reject loan requests from a queue of applicants from either of two groups with ID 1 or 2 respectively. The applicants arrive one-by-one in a sequential manner. At each time step t, the applicant's group ID g_t is sampled uniformly from $G = \{1, 2\}$. Given the current applicant's group ID $g_t \in \{0, 1\}$, the corresponding credit score $c_t \in \{1, 2, \cdots, C\}$ is sampled from the credit distribution $\boldsymbol{\mu}_{t,g_t} \in \Delta(C)$, where $\Delta(C)$ denotes the set of all discrete distributions over $\{1, 2, \cdots, C\}$. We note here that the credit distributions of both groups, $\boldsymbol{\mu}_{t,1}$ and $\boldsymbol{\mu}_{t,2}$ are time-varying and will introduce their dynamics in detail later. Regardless of their group IDs g_t, the applicants with higher credit score is more likely to repay (i.e., $Y_t = 1$), whether the loan is approved (i.e., $\hat{Y}_t = 1$) or not (i.e., $\hat{Y}_t = 0$). Group 2 is disadvantaged with a lower mean of initial credit score compared to Group 1 in the beginning of the sequential decision-making process. The agent makes the decision $\hat{Y}_t \in \{0, 1\}$ using the observation g_t and c_t. With $\hat{Y}_t$ and Y_t, the agent gets an immediate reward R_t (agent's earned cash at step t), and the credit score distribution of group G_t changes depending on $\hat{Y}_t$ and Y_t.

While trying to maximizing the cumulative reward, the agent also tries to balance the group well-being measured by the true positive rate. The fairness criterion here is an long-term extension of EO defined as

$$\left| \frac{\sum_t \mathbb{1}\{G_t = 0, Y_t = \hat{Y}_t = 1\}}{\sum \mathbb{1}\{G_t = 0, Y_t = 1\}} - \frac{\sum_t \mathbb{1}\{G_t = 1, Y_t = \hat{Y}_t = 1\}}{\sum \mathbb{1}\{G_t = 1, Y_t = 1\}} \right| \tag{11}$$

As for the dynamics of both group's credit score distribution, given the current group ID g_t, the credit score will shift according to $\hat{Y}_t$ and Y_t. Specifically, the credit score of current applicant will be affected and shift form c_t to a new score c_t', i.e.,

$$\boldsymbol{\mu}_{t+1,g_t}(c_t') - \boldsymbol{\mu}_{t,g_t}(c_t') = \boldsymbol{\mu}_{t,g_t}(c_t) - \boldsymbol{\mu}_{t+1,g_t}(c_t) = \varepsilon \geq 0. \tag{12}$$

In the original setting from (Liu et al., 2018), the credit score will shift deterministically only if the loan is approved. It will increase by 1 if the loan is repaid, otherwise it will decrease by 1. We use the rows of right stochastic matrices $\mathbf{P}_{g_t,\hat{Y},Y} \in \mathbb{R}^{C \times C}$ to represent the distribution of c_t' given c_t with the specific g_t, Y and $\hat{Y}$. In the original setting, $C = 7, \varepsilon = 0.01$ and $\{\mathbf{P}_{g_t,\hat{Y},Y}\}$ are given by $\mathbf{P}_{g_t,0,0} = \mathbf{P}_{g_t,0,1} = \mathbf{I}_{7\times 7}$ and

$$\mathbf{P}_{g_t,1,0} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{P}_{g_t,1,1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

for all $g_t \in \{1, 2\}$. The initial credit score distributions of two groups are given by

$$\boldsymbol{\mu}_{0,1} = \begin{bmatrix} 0 & 0.1 & 0.1 & 0.2 & 0.3 & 0.3 & 0 \end{bmatrix}, \boldsymbol{\mu}_{0,2} = \begin{bmatrix} 0.1 & 0.1 & 0.2 & 0.3 & 0.3 & 0 & 0 \end{bmatrix}. \tag{14}$$

In addition, we extend the original setting into a more general case. The first modification is that the credit score of the rejected applicant from Group $g_t \in \{1, 2\}$ still shift as if the loan was accepted with probability $\delta_{g_t} > 0$. The seconds difference is that the credit score shifts in a stochastic manner rather than in a deterministic way as before. Specifically, we keep $C = 7, \varepsilon = 0.01$ and set $\delta_1 = 0.3, \delta_2 = 0.1$. The modified $\{\mathbf{P}_{g_t,\hat{Y},Y}\}$ are given by

$$\mathbf{P}_{g_t,1,0} = \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0.9 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.2 & 0.7 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.7 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.7 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0.7 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.1 \end{bmatrix}, \ \mathbf{P}_{g_t,1,1} = \begin{bmatrix} 0.2 & 0.7 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.7 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.7 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0.7 & 0.1 & 0 \\ 0 & 0.5 & 0 & 0 & 0.2 & 0.2 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

and $\mathbf{P}_{g_t,0,Y} = \delta_{g_t}\mathbf{P}_{g_t,1,Y} + (1-\delta_{g_t})\mathbf{I}_{7\times 7}$ for all $g_t \in \{1,2\}$ and $Y \in \{0,1\}$. We modify the initial credit score distributions of two groups to

$$\boldsymbol{\mu}_{0,1} = \begin{bmatrix} 0.05 & 0.05 & 0 & 0.2 & 0.1 & 0.3 & 0.3 \end{bmatrix}, \ \boldsymbol{\mu}_{0,2} = \begin{bmatrix} 0 & 0 & 0.3 & 0.3 & 0.3 & 0.05 & 0.05 \end{bmatrix}. \tag{16}$$

For both the original setting and our new setting, the repaying probability of applicant with credit score g is given by

$$\{\rho_g\}_{g=1}^{7} = \{0.1,\ 0.2,\ 0.45,\ 0.6,\ 0.65,\ 0.7,\ 0.7\}. \tag{17}$$

Such modification introduces more stochasticity into the environment, which requires the RL agent to account for more long-term effects and thus is more challenging. As shown in Section 5, the proposed ELBERT-PO obtains high utility and low bias in such challenging environment.

Attention allocation. In the third environment, the agent's task is to allocate 6 attention units to 5 sites to discover incidents, where each site has different initial incident rate. Since each site is considered a group, this environment is in a multi-group setting. To describe the dynamics, let $a_{g,t}$ and $\mu_{g,t}$ be the allocated attention and incident rate for the group g at time t. The number of incidents $y_{g,t}$ is sampled from $\text{Poisson}(\mu_{g,t})$ with incident rate $\mu_{g,t}$, and the number of discovered incident is $\hat{y}_{g,t} = \min(a_{g,t}, y_{g,t})$. The incident rate changes according to $\mu_{g,t+1} = \mu_{g,t} - d \cdot a_{g,t}$ if $a_{g,t} > 0$ and $\mu_{g,t+1} = \mu_{g,t} + d$ otherwise, where d is a constant. The agent's reward is $R(s_t, a_t) = \sum_g \hat{y}_{g,t} - \zeta \sum_g (y_{g,t} - \hat{y}_{g,t})$, where the coefficient ζ balances between discovering and missing incidents.

Here the group well-being is defined as the ratio between the total number of discovered incidents over time and the total number of incidents, and thus the bias is defined as

$$\max_{g\in G} \frac{\sum_t \hat{y}_{g,t}}{\sum_t y_{g,t}} - \min_{g\in G} \frac{\sum_t \hat{y}_{g,t}}{\sum_t y_{g,t}}. \tag{18}$$

For the parameter, we keep $\zeta = 0.25$ and use $d = 0.1$ as the dynamic rate. The initial incident rate is given by

$$\{\mu_{g,0}\}_{g=1}^{5} = \{8,\ 6,\ 4,\ 3,\ 1.5\} \tag{19}$$

as same as the original setting.

Attention allocation: original version. In the original version of this environment used in (Yu et al., 2022), the agent's task is to allocate 6 attention units to 5 sites (groups) to discover incidents, where each site has a different initial incident rate. The agent's action is $a_t = \{a_{g,t}\}_{g=1}^{5}$, where $a_{g,t}$ is the number of allocated attention units for group g. The number of incidents $y_{g,t}$ is sampled from $\text{Poisson}(\mu_{g,t})$ with incident rate $\mu_{g,t}$ and the number of discovered incident is $\hat{y}_{g,t} = \min(a_{g,t}, y_{g,t})$. The incident rate changes according to $\mu_{g,t+1} = \mu_{g,t} - d \cdot a_{g,t}$ if $a_{g,t} > 0$ and $\mu_{g,t+1} = \mu_{g,t} + d$ otherwise, where the dynamic rate d is a constant. The agent's reward is $R(s_t, a_t) = \sum_g \hat{y}_{g,t} - \zeta \sum_g (y_{g,t} - \hat{y}_{g,t})$, where the coefficient ζ balances between the discovered and missed incidents. In the original version, $\zeta = 0.25$ and $d = 0.1$. The initial incident rates are given by

$$\{\mu_{g,0}\}_{g=1}^{5} = \{8,\ 6,\ 4,\ 3,\ 1.5\}. \tag{20}$$

The group well-being is defined as the ratio between the total number of discovered incidents over time and the total number of incidents, and thus the bias is defined as

$$\max_{g\in G} \frac{\sum_t \hat{y}_{g,t}}{\sum_t y_{g,t}} - \min_{g\in G} \frac{\sum_t \hat{y}_{g,t}}{\sum_t y_{g,t}}. \tag{21}$$

Attention allocation: harder version. To modify the environment to be more challenging, we consider a more general environment by introducing more complexity. Different from the original setting in (Yu et al., 2022) where the dynamic rate is the same among groups, we consider a more general case where the dynamic rates vary among different groups. Moreover, for the group g, the dynamic rate for increasing incident rate $\overline{d}_g$ is different from that for decreasing incidient rate $\underline{d}_g$. Specifically, the incident rate changes according to $\mu_{g,t+1} = \mu_{g,t} - \underline{d}_g \cdot a_{g,t}$ if $a_{g,t} > 0$ and $\mu_{g,t+1} = \mu_{g,t} + \overline{d}_g$ otherwise, where the constants $\underline{d}_g$ and $\overline{d}_g$ are the dynamic rates for reduction and growth of the incident rate of group g. The parameters are given by the following.

$$\{\underline{d}_g\}_{g=1}^5 = \{0.004,\ 0.01,\ 0.016,\ 0.02,\ 0.04\},\ \{\overline{d}_g\}_{g=1}^5 = \{0.08,\ 0.2,\ 0.4,\ 0.8,\ 2\} \tag{22}$$

Meanwhile, we increase the number of attention units to allocate from 6 to 30 and the initial incident rates to

$$\{\mu_{g,0}\}_{g=1}^5 = \{30,\ 25,\ 22.5,\ 17.5,\ 12.5\}. \tag{23}$$

The agent's reward is $R(s_t, a_t) = -\zeta \sum_g (y_{g,t} - \hat{y}_{g,t})$, i.e., the opposite of the sum of missed incidents. Here $\zeta = 0.25$. Note that we use a different reward function from the original setting.

Explanation of the harder environment. The new version of the attention environment is more challenging for learning a fair policy with high rewards due to the following reasons. **(1)** The higher number of attention units indicates the larger action space in which searching for the optimal policy will be more challenging. **(2)** For all groups, the increasing dynamic rates are much higher than the decreasing dynamic rates, making it harder for the incident rate to decrease. **(3)** The disadvantaged groups, i.e., the groups with higher initial incident rates, have lower dynamic rates for both decreasing and increasing incident rate. This makes learning a fair policy harder since lower decreasing dynamic rates make the incident rates harder to decrease, and lower increasing dynamic rates means the policy could allocate less units to these groups without harming the reward too much, causing increasing bias.

C.2. Hyperparameters

For each method in all three environments, we use 10^{-6} as the learning rate in lending and 10^{-5} in attention environment, and train for 5×10^6 time steps.

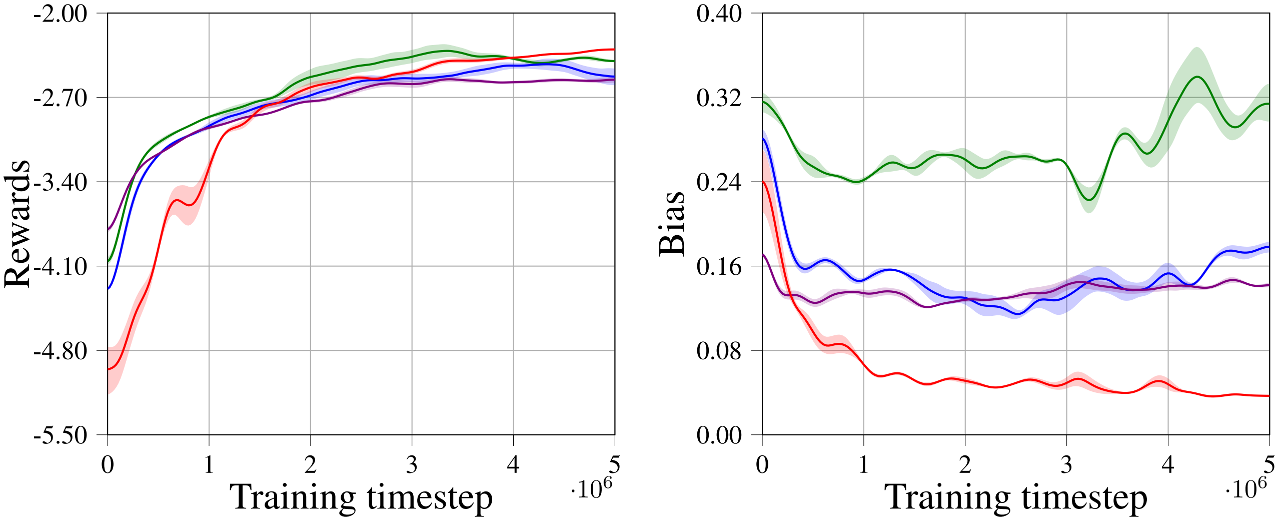
For the bias coefficient α for ELBERT-PO, we use $\alpha = 400$ in lending and $\alpha = 20000$ in other two environments. In attention allocation with multiple groups, we set the temperature β of soft bias as 20. For the hyperparameters of baseline method R-PPO, we choose $\zeta_0 = 1$ in all environments, and $\zeta_1 = 2$ in lending, $\zeta_1 = 10$ in attention allocation. For the hyperparameters of baseline method A-PPO, we choose $\beta_0 = 1$ in all environments, and $\beta_1 = \beta_2 = 0.25,\ \omega = 0.005$ in lending, $\beta_1 = \beta_2 = 0.15,\ \omega = 0.05$ in attention allocation.

All experiments are run on NVIDIA GeForce RTX 2080 Ti GPU.

C.3. Experiments in multi-group settings

In this section, we demonstrate the preliminary results of ELBERT-PO in the multi-group setting using the the harder version of attention allocation environment (Atwood et al., 2019).

Results: ELBERT-PO reduces the bias and obtains high reward in the multi-group setting. ELBERT-PO achieves the lowest bias and the highest reward of among all methods. This shows the effectiveness of ELBERT-PO in the multi-group setting.



(a) Rewards

(b) Bias

Figure 5. Learning curve for the harder attention allocation environment.