Reinforcement Learning with Stepwise Fairness Constraints

Zhun Deng
Columbia University

He Sun
Harvard University

Zhiwei Steven Wu
Carnegie Mellon University

Linjun Zhang
Rutgers University

David C. Parkes
DeepMind
Harvard University

Abstract

AI methods are used in societally important settings, ranging from credit to employment to housing, and it is crucial to provide fairness in regard to automated decision making. Moreover, many settings are dynamic, with populations responding to sequential decision policies. We introduce the study of reinforcement learning (RL) with stepwise fairness constraints, which require group fairness at each time step. In the case of tabular episodic RL, we provide learning algorithms with strong theoretical guarantees in regard to policy optimality and fairness violations. Our framework provides tools to study the impact of fairness constraints in sequential settings and brings up new challenges in RL.

1 INTRODUCTION

Automated decision making systems are increasingly used in our daily lives, for example, in the context of lending, insurance, and medical care. A challenge is that these decision systems may demonstrate discrimination against disadvantaged groups (Dwork et al., 2012). In order to mitigate this issue, fairness constraints have been proposed (Hardt et al., 2016; Dwork et al., 2012), for example looking to achieve certain statistical parity properties. Despite the fact that fair machine learning has been extensively studied, most of this work considers the static setting without considering the sequential feedback effects of decisions. At the same time, algorithmic decisions may lead to changes in the underlying statistical patterns in data, through feedback loops with society. In turn, this affects the decision making process; for example, the decisions of banks may lead borrowers to react, perhaps closing credit cards and changing their FICO scores.

When there exist sequential feedback effects, even ignoring one-step feedback effects can harm minority groups (Liu et al., 2018). In response, Zhang et al. (2020) study a discrete-time sequential decision process, where responses to the decisions made at each time step are accompanied by changes in the features and qualifications of the population in the next time step. In particular, they study and show the drawback of myopic optimization together with requiring fairness at each time step, i.e., *stepwise fairness constraints*. Imposing stepwise fairness is a natural way to incorporate fairness into a Markov decision process (MDP). At the same time, it is sensible to consider fairness alongside considerations of long-term reward. In this paper, we take the perspective of a forward-looking decision maker, combining stepwise fairness with optimal, sequential-decision making.

We initiate both the theoretical and experimental studies of reinforcement learning under stepwise fairness constraints. Our work can be viewed as a *Fair Partially Observable Markov Decision Process* (F-POMDP) framework to promote fair sequential decision making. Our work also provides a computational tool for studying the quantities of interests, especially the well-being of different groups, in a natural sequential decision making setting.

We consider an episodic setting, which models for example economic and societal activities that exhibit seasonality; e.g., new mortgage applicants who apply for loans from banks more often in the spring and summer season every year, or graduate school admission, which usually starts in the autumn and completes around December every year. Similar to Liu et al. (2018) and Zhang et al. (2020), we mainly consider two types of fairness notions, those of *demographic parity* and *equalized opportunity*. These are illustrative of other stepwise fairness constraints that could be adopted. We adopt a POMDP framework that has discrete actions and a discrete state

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.
Emails: zhundeng@g.harvard.edu, he_sun@g.harvard.edu

space and take a model-based learning approach, giving practical optimization algorithms that enjoy strong theoretical guarantees in regard to policy optimality and fairness violations as the number of episodes increases. We summarize our contributions as below:

1. Theoretically, we demonstrate how to use sampled trajectories of individuals to solve RL with fairness constraints, and provide theoretical guarantees to ensure vanishing regrets in reward and fairness violation as the number of episodes increases.

2. Experimentally, we implement and evaluate the first algorithm for tabular episodic RL with stepwise fairness constraints.

1.1 Related Work

There is increasing interest in the study of decision making problems in the context of people (Hardt et al., 2015; Shavit and Moses, 2019; Ball, 2019; Chen et al., 2020). Hardt et al. (2015) model a classification problem as a sequential game (Stackelberg competition) between two players, where the first player has the ability to commit to his strategy before the second player responds. They characterize the equilibruim and obtain near optimal computationally efficient learning algorithms. Shavit and Moses (2019) study an algorithmic decision-maker who incentivizes people to act in certain ways to receive a better decision. Ball (2019) studies a model of predictive scoring, where there is a sender agent being scored, a receiver agent who wants to predict the quality of the sender, and an intermediary who observes multiple, potentially mutable features of the sender.

There is also a growing literature on algorithmic fairness (Liu et al., 2018; Calders et al., 2009; Kusner et al., 2017; Dwork et al., 2012; Burhanpurkar et al., 2021; Deng et al., 2022, 2023). Liu et al. (2018), for example, characterize the delayed impact of standard fairness criteria under a feedback model with a single period of adaptation. They use a one step feedback model to capture the sequential dynamics of the environment. However, these papers do not consider fairness in a more general sequential decision process. There is also a line of literature regarding the special case of fair bandits (Joseph et al., 2016; Hashimoto et al., 2018).

In regard to fairness considerations in reinforcement learning, this is also gaining recent attention (D'Amour et al., 2020; Creager et al., 2019; Wen et al., 2021; Jabbari et al., 2017; Mandal and Gan, 2022). In particular, Creager et al. (2019) use causal directed acyclic graphs as a unifying framework for fairness. D'Amour et al. (2020) use simulation to study the fairness of algorithms and show that neither static nor single-step analyses is enough to understand the long-term consequences of a decision system. Jabbari et al. (2017) define fairness constraints to require that an algorithm never prefers one action over another if the long-term reward of choosing the latter action is higher, whereas we consider group-wise notions of fairness. Mandal and Gan (2022) adopt a welfare-based, axiomatic approach, and give a regret bound for the Nash Social, Minimum and generalized Gini Welfare. In contrast with our work, their fairness concepts are not group-based but rather based on the value contributed from different agents in the system. Zhang et al. (2020) also study the dynamics of population qualification and algorithmic decisions under a POMDP problem setting, but whereas they only consider myopic policies we formulate this as a general reinforcement learning policies.

2 PRELIMINARIES

We consider a binary decision setting, with training examples that consist of triplets (x, y, ϑ), where $x \in \mathcal{X}$ is a feature vector, $\vartheta \in \Lambda$ is a protected group attribute such as race or gender, and the label $y \in \{0, 1\}$. For simplicity, we only consider binary sensitive attributes $\Lambda = \{\alpha, \beta\}$, but our method can also be generalized to deal with multiple sensitive attributes (see Appendix B). For $k \in \mathbb{N}^+$, we use $[k]$ to denote the set $\{1, 2, \cdots, k\}$.

Based on feature x, a decision maker makes a decision $a \in \mathcal{A} = \{0, 1\}$ (e.g., make a loan or not). We also denote an individual's state as $s = (x, y)$, and let $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$. After a decision is made, a possibly group-dependent reward, which may be stochastic, $r^\vartheta : (s, a) \mapsto \mathbb{R}$, is obtained by the decision maker. A concrete example of a reward function, with $r_+, r_- > 0$, is

$$r^\vartheta(s, a) = \begin{cases} r_+, & \text{if } y = 1, a = 1; \\ -r_-, & \text{if } y = 0, a = 1; \\ 0, & \text{if } a = 0. \end{cases}$$

Here, the decision maker gains r_+ by accepting a qualified individual and incurs a cost r_- by accepting an unqualified individual.

2.1 Sequential Setting

Our model as a partially observable Markov decision process (POMDP) mainly follows Zhang et al. (2020), but whereas they consider a fair, myopically-optimizing policy we consider long-term rewards as in typical RL settings. Following Liu et al. (2018) and Zhang et al. (2020), the decision maker is interested in the expected reward achieved across time for a ***random individual*** drawn from the population. Each random individual

has their group membership sampled according to $p_\alpha = \mathbb{P}(\vartheta = \alpha)$ and $p_\beta = \mathbb{P}(\vartheta = \beta)$, and interacts with the decision maker over multiple time steps. At each time step h, the sampled individual with attribute ϑ has feature $x_h^\vartheta = x^\vartheta \in \mathcal{X}$ along with a hidden qualification status $y_h^\vartheta = y^\vartheta \in \{0, 1\}$. An example is that the feature x_h^ϑ may be determined by the hidden qualification status, with $x_h^\vartheta \sim p(\cdot|y^\vartheta)$. We call $s_h^\vartheta = (x_h^\vartheta, y_h^\vartheta)$ the *state* of the individual at time h. The initial state s_1^ϑ is sampled from p^ϑ.

At each time step h, the decision maker adopts a decision a_h^ϑ based on the observed feature x^ϑ by following a *group-dependent policy* $\pi_h^\vartheta(x^\vartheta)$, i.e. $a_h^\vartheta \sim \pi_h^\vartheta(x_h^\vartheta)$, where $\pi_h^\vartheta : \mathcal{X} \to \Delta(\mathcal{A})$, and $\Delta(\mathcal{A})$ is the set of distributions on $\mathcal{A}$.[1] The decision maker receives *reward* $r^\vartheta(s_h^\vartheta, a_h^\vartheta)$, and $r^\vartheta \in [l, u]$, where $-\infty < l - u < \infty$. Without loss of generality, we assume $r^\vartheta \in [0, 1]$. The mean of stochastic reward function is denoted by $r^{*\vartheta}$.

After the decision is made, the individual is informed of the decision and their qualification status, y_{h+1}^ϑ, and features x_{h+1}^ϑ, may then undergo a stochastic transition. We assume that this transition follows a *time-invariant* but *group-dependent transition kernel*, which we denote as $p^{*\vartheta}(s_{h+1}^\vartheta|s_h^\vartheta, a_h^\vartheta)$, where $p^{*\vartheta} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, and $\Delta(\mathcal{S})$ is the set of distributions on $\mathcal{S}$. As explained in Zhang et al. (2020), in addition to thinking about a single, randomly chosen individual repeatedly interacting with the decision maker, this also models a finite population of randomly chosen individuals, some from each group, and with all individuals in a group subject to the same, group-contingent decision policies. In addition, we have a reward function pair, (r^α, r^β), which may be stochastic.

2.2 Fair Policies

The goal of the decision maker is to find a *policy*, $\boldsymbol{\pi} = (\pi^\alpha, \pi^\beta)$, to maximize the total, expected reward over an episode for a random individual, while satisfying stepwise fairness constraints, i.e. imposing a certain type of fairness constraint on states and actions at each time step.

A random individual in group ϑ, and with decision policy π^ϑ, follows stochastic state, action, reward sequence $s_1^\vartheta, a_1^\vartheta, r^\vartheta(s_1^\vartheta, a_1^\vartheta), s_2^\vartheta, a_2^\vartheta, r^\vartheta(s_2^\vartheta, a_2^\vartheta), s_3^\vartheta, \cdots$. Let $\mathbb{E}^{\pi,p}[\cdot], \mathbb{P}^{\pi,p}[\cdot]$ be the expectation and probability of a random variable defined with respect to this stochastic process. We denote the expected reward for a random individual at time step h as,

$$\mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi}) = \sum_{\vartheta \in \{\alpha,\beta\}} p_\vartheta \cdot \mathbb{E}^{\pi^\vartheta, p^{*\vartheta}}[r^{*\vartheta}(s_h^\vartheta, a_h^\vartheta)].$$

Here, $\mathbb{E}^{\pi^\vartheta, p^{*\vartheta}}[\cdot]$ refers to the expected value for an individual in group ϑ, i.e., conditioned on the individual being sampled in this group. Our goal is to obtain the policy $\boldsymbol{\pi}^*$ that solves the following optimization problem

$$\max_{\boldsymbol{\pi}} \sum_{h=1}^{H} \mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi})$$
$$\text{s.t. } \forall h \in [H],$$
$$faircon(\{\pi^\vartheta, p^{*\vartheta}, s_h^\vartheta, a_h^\vartheta\}_{\vartheta=\{\alpha,\beta\}}),$$

where *faircon* corresponds to a particular fairness concept. We consider two fairness concepts, and discuss how the approach can be extended to other concepts (see Section 6.2). Specifically, we consider the following two group fairness concepts:

(i) *RL with demographic parity (DP).* For this case, *faircon* is

$$\mathbb{P}^{\pi^\alpha, p^{*\alpha}}[a_h^\alpha = 1] = \mathbb{P}^{\pi^\beta, p^{*\beta}}[a_h^\beta = 1],$$

and at each time step h, the decision for individuals from different groups is statistically independent of the sensitive attribute.

(ii) *RL with equalized opportunity (EqOpt).* For this case, *faircon* is

$$\mathbb{P}^{\pi^\alpha, p^{*\alpha}}[a_h^\alpha = 1 | y_h^\alpha = 1] = \mathbb{P}^{\pi^\beta, p^{*\beta}}[a_h^\beta = 1 | y_h^\beta = 1],$$

and at each time step h, the decision for a random individual from each of the two different groups is statistically independent of the sensitive attribute conditioned on the individual being qualified.

In each case, $\mathbb{P}^{\pi^\alpha, p^{*\alpha}}[\cdot]$ refers to the probability for an individual in group ϑ, i.e., conditioned on the individual being sampled in this group.

The above optimization problems are feasible under technical Assumption 4.1 (presented later), and we assume feasibility throughout the paper (see also Appendix C).

Remark 2.1. *Since we are modelling the behavior of a randomly drawn individual from the population, the objective should be viewed as to find a policy pair $\boldsymbol{\pi} = (\pi^\alpha, \pi^\beta)$ to optimize the long-term reward of of the decision maker while ensuring fairness over the population.*

[1] Following Zhang et al. (2020), we use group-dependent policies so that the formulation can be more generalized. Our technique can also be used for group-independent policies if this is required.

2.3 Episodic RL Protocol

This is a learning setting, and we study an episodic sequential decision setting where a learner repeatedly interacts with an environment across $K > 0$ independent episodes. Such a scenario is natural in a number of practical settings, as we discussed in the introduction. We consider the tabular case, i.e., we assume finite cardinality for $\mathcal{S}$ and $\mathcal{A}$.[2] Without loss of generality, we assume that the initial state of an episode is a fixed state s_0 (the next state can be sampled randomly, and state s_0 does not contribute to any reward or fairness considerations, see Brantley et al. (2020) for further detailed explanation). At episode $k \in [K]$, denote policy pair $\boldsymbol{\pi}_k = (\pi_k^\alpha, \pi_k^\beta) = \{(\pi_{k,h}^\alpha, \pi_{k,h}^\beta)\}_{h=1}^H$, where H is the horizon. An individual sampled from group G_ϑ starts from state $s_{k,1}^\vartheta$, thus, we can consider starting state pair $(s_{k,1}^\alpha, s_{k,1}^\beta) = (s_0^\alpha, s_0^\beta)$ for the trajectory of different groups (the initial state depends on the group from which the individual is drawn). At each time step $h \in [H]$, the decision maker selects an action $a_{k,h}^\vartheta \sim \pi_{k,h}^\vartheta(x_{k,h}^\vartheta)$. Here, although the policy only uses the x component, it is convenient to write it as a function of s. The decision maker gets reward $r^\vartheta(s_{k,h}^\vartheta, a_{k,h}^\vartheta)$, and the state of the individual for next time step is drawn according to $s_{k,h+1}^\vartheta \sim p^{*\vartheta}(\cdot|s_{k,h}^\vartheta, a_{k,h}^\vartheta)$.

Remark 2.2. *The policy is only based on the **feature vector** x. However, we are able to access y in the training data.*

3 LEARNING ALGORITHMS

Before we formally state our algorithms, we need to introduce the data used to estimate the unknown quantities such as reward function mean pair $\boldsymbol{r}^* = (r^{*\alpha}, r^{*\beta})$ and transition probability pair $\boldsymbol{p}^*$. In addition, we will incorporate an *exploration bonus* to further modify the estimation.

Data gathering and estimation. In order to analyze the policy effect on the population, we model the behavior of a randomly drawn individual who interacts with the environment across H steps. Here, we demonstrate how to aggregate individuals' data for each episode and estimate quantities of interest. Specifically, at episode $k \in [K]$, for each group ϑ, we assume n_k^ϑ individuals are drawn, according to p_α and p_β. Throughout the paper, we assume $n_k^\vartheta \geq 1$, for each ϑ and k.

A decision is made independently for each individual at each step, using a group-specific policy, leading to a stochastic transition in the state of the individual. In Appendix D, we will further discuss how to gather data when allowing individuals who opt in or out during an episode. We use the counting method to obtain estimates of the statistics of interest. For the i-th individual in episode k, their status and action at time step h is denoted as $s_{k,h}^{\vartheta,i}$ and $a_{k,h}^{\vartheta,i}$.

For $\vartheta \in \{\alpha, \beta\}$, let $\boldsymbol{p}_k = (p_k^\alpha, p_k^\beta)$ and $\boldsymbol{r}_k = (r_k^\alpha, r_k^\beta)$,

$$N_k^\vartheta(s,a) = \max\Big\{1, \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} \mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a\}\Big\},$$

$$p_k^\vartheta(s'|s,a) = \frac{1}{N_k^\vartheta(s,a)} \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} \mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a, s_{k,h+1}^{\vartheta,i} = s'\},$$

$$\hat{r}_k^\vartheta(s,a) = \frac{1}{N_k^\vartheta(s,a)} \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} r^\vartheta(s,a)\mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a\}.$$

Exploration bonus method. In RL, it is standard to introduce optimism in order to encourage exploring under-explored states. Specifically, for $\vartheta \in \{\alpha, \beta\}$, we adopt a *bonus term*, $\hat{b}_k^\vartheta$, to add to the estimated reward function $\hat{r}_k^\vartheta$, such that we obtain $r_k^\vartheta(s,a) = \hat{r}_k^\vartheta(s,a) + \hat{b}_k^\vartheta(s,a)$, where the $\hat{b}_k^\vartheta(s,a)$ values assign larger values for under-explored (s,a)'s. We specify how to choose $\hat{b}_k^\vartheta$ in Section 4.1, and denote

$$\mathcal{R}_{k,h}(\boldsymbol{p}, \boldsymbol{\pi}) = \sum_{\vartheta\in\{\alpha,\beta\}} p_\vartheta \cdot \mathbb{E}^{\pi^\vartheta, p^\vartheta}[r_k^\vartheta(s_{k,h}^\vartheta, a_{k,h}^\vartheta)].$$

For the purpose of analysis, we treat p_ϑ's as known constants for simplicity; for example, perhaps these proportions are provided by census.

Practical optimization for DP. In reality, given we don't have access to $\boldsymbol{p}^*$ and $r^{*\vartheta}$, we need to solving a surrogate optimization problem and hope the optimal policy can have similar performance as the ideal optimal policy under certain performance criteria. In the following, we provide a simple algorithm for RL with demographic parity. It is based on optimization under p_k^ϑ and r_k^ϑ:

$$\max_{\boldsymbol{\pi}\in\Pi_k} \sum_{h=1}^H \mathcal{R}_{k,h}(\boldsymbol{p}_k, \boldsymbol{\pi}),$$
$$s.t.\ \forall h \in [H],$$
$$|\mathbb{P}^{\pi^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi^\beta, p_k^\beta}(a_{k,h}^\beta = 1)| \leq \hat{c}_{k,h},$$

[2] In the example of credit score and loan payment in Liu et al. (2018), the credit score is discretized and served as $\mathcal{X}$ here and the action space $\mathcal{A}$ and qualification status space $\mathcal{Y}$ are both $\{0,1\}$.

where we have $\Pi_k = \{(\pi^\alpha, \pi^\beta) : \pi^\vartheta(a = 1|x) \geq \eta_k^{DP}, \forall x, a, h, \vartheta\}$. Π_k can ensure the reachability from any x to the decision $a = 1$. Here $\{\eta_k^{DP}\}_k$ is a sequence of real numbers and $\{\hat{c}_{k,h}\}_{k,h}$ are relaxations. Intuitively, if we set η_k^{DP} and $\hat{c}_{k,h}$ to be decreasing and vanishing as k increases, we would expect the above optimization to approach the ideal optimization problem as k increases. We formalize this in Section 4.1.

Practical optimization for EqOpt. For equalized opportunity, and similar to the case of DP, we have

$$\max_{\pi \in \Pi_k} \sum_{h=1}^{H} \mathcal{R}_{k,h}(\boldsymbol{p}_k, \boldsymbol{\pi}),$$
$$s.t.\ \forall h \in [H],\ \ |\mathbb{P}^{\pi^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1 | y_{k,h}^\alpha = 1) - \mathbb{P}^{\pi^\beta, p_k^\beta}(a_{k,h}^\beta = 1 | y_{k,h}^\beta = 1)| \leq \hat{d}_{k,h},$$

where we have $\Pi_k = \{(\pi^\alpha, \pi^\beta) : \pi^\vartheta(a = 1|x) \geq \eta_k^{EqOpt}, \forall x, a, h, \vartheta\}$. Here $\{\eta_k^{EqOpt}\}_k$ is a sequence of real numbers and $\{\hat{d}_{k,h}\}_{k,h}$ are relaxations. We formalize this in Section 4.1.

Algorithm. We can solve these optimization problems through occupancy measures, and they each become different kinds of quadratically constrained linear programs (QCLP) (see Appendix A). Although QCLP is generally NP-hard, many methods based on relaxations and approximations such as semi-definite program (SDP) have been extensively discussed. We use Gurobi to solve these relaxed optimization problems.

4 THEORETICAL ANALYSIS

In order to track the performance of the algorithm, we consider the following regrets. For policy pairs $\{\boldsymbol{\pi}_k\}_{k=1}^K$, for *reward regret* in episode k, we track:

$$\mathcal{R}_{\text{reg}}^{\text{type}}(k) = \frac{1}{H} \sum_{h=1}^{H} \Big(\mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi}^{*\text{type}}) - \sum_{t=1}^{k} \mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi}_k) \Big),$$

where $\boldsymbol{\pi}^{*\text{type}}$ is the optimal policy pair of RL with constraint types mentioned above and type $\in \{DP, EqOpt\}$. For simplicity, we will omit the superscript "type" when it is clear from the context and use $\boldsymbol{\pi}^* = (\pi^{*\alpha}, \pi^{*\beta})$.

For the fairness constraints, we consider the violation for each type of constraint in episode k as the following,

$$\mathcal{C}_{\text{reg}}^{DP}(k) = \frac{1}{H} \sum_{h=1}^{H} \left| \mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1) \right|,$$

and

$$\mathcal{C}_{\text{reg}}^{EqOpt}(k) = \frac{1}{H} \sum_{h=1}^{H} \left| \mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1 | y_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1 | y_{k,h}^\beta = 1) \right|.$$

The theoretical guarantees hold for any episode, not just the last episode.

4.1 Choice of Various of Quantities

In this part, we provide a formal theoretical guarantee for the performance of our algorithms under the previously mentioned criteria with suitably chosen quantities $\{\hat{b}_k^\vartheta\}_k$, $\{\hat{c}_k\}_k$, and $\{\hat{d}_k\}_k$, for each episode k.

Q and V functions. Two of the mostly used concepts in RL are Q and V functions. Specifically, Q functions track the expected reward when a learner starts from state $s \in \mathcal{S}$. Meanwhile, V functions are the corresponding expected Q functions of the selected action. For a reward function r and a transition function p, the Q and V functions are defined as:

$$Q_r^{\pi,p}(s, a, h) = r(s, a) + \sum_{s' \in \mathcal{S}} p(s'|s, a) V_r^{\pi,p}(s', h + 1),$$
$$V_r^{\pi,p}(s, h) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q_r^{\pi,p}(s, a, h)],$$

where we set $V_r^{\pi,p}(s, H + 1) = 0$.

Choice of $\hat{b}_k^\vartheta$. For $\{\hat{b}_k^\vartheta\}_k$, similar to Brantley et al. (2020), we need $\{\hat{b}_k^\vartheta\}_k$ to be valid.

Definition 4.1 (Validity)**.** *A bonus $\hat{b}_k^\vartheta$ is valid for episode k if for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,*

$$\Big| \hat{r}_k^\vartheta(s, a) - r_k^{*\vartheta}(s, a) + \sum_{s' \in \mathcal{S}} \Big(p_k^\vartheta(s'|s, a) - p^{*\vartheta}(s'|s, a) \Big) V_{r^{*\vartheta}}^{\pi^{*\vartheta}, p^{*\vartheta}}(s', h + 1) \Big| \leq \hat{b}_k^\vartheta(s, a).$$

Following the exporation-bonus setting (Brantley et al., 2020), we set $\hat{b}_k^\vartheta = \min\Big\{2H, 2H\sqrt{\frac{2\ln(16SAHk^2/\delta)}{N_k^\vartheta(s,a)}}\Big\}$, and have the following lemma.

Lemma 4.1. *With probability at least $1 - \delta$, for*

$$\hat{b}_k^\vartheta(s, a) = \min\Big\{2H, 2H\sqrt{\frac{2\ln(16SAHk^2/\delta)}{N_k^\vartheta(s, a)}}\Big\},$$

the bonus $\hat{b}_k^\vartheta(s, a)$ is valid for every episode k simultaneously.

Choice of $\hat{c}_{k,h}$ and $\hat{d}_{k,h}$. For $\{\hat{c}_{k,h}\}_{k,h}$ and $\{\hat{d}_{k,h}\}_{k,h}$, we require them to be *compatible*.

Definition 4.2 (Compatibility). *The sequence $\{\hat{c}_{k,h}\}_{k,h}$ is* compatible *if for all $h \in [H], k \in [K]$, $|\mathbb{P}^{\pi^{*\alpha},p_k^{\alpha}}(a_h = 1) - \mathbb{P}^{\pi^{*\beta},p_k^{\beta}}(a_h = 1)| \leq \hat{c}_{k,h}$. The sequence $\{\hat{d}_{k,h}\}_{k,h}$ is* compatible *if for all $h \in [H], k \in [K]$, $|\mathbb{P}^{\pi^{*\alpha},p_k^{\alpha}}(a_h = 1|y_h = 1) - \mathbb{P}^{\pi^{*\beta},p_k^{\beta}}(a_h = 1|y_h = 1)| \leq \hat{d}_{k,h}$*

Briefly speaking, we hope that when substituting $p^{*\vartheta}$ to p_k^{ϑ}, that $\hat{c}_{k,h}$ and $\hat{d}_{k,h}$ can control the fairness constraints violation. Let us use S to denote $|\mathcal{S}|$ and A to denote $|\mathcal{A}|$.

Lemma 4.2. *Denote $N_k^{\vartheta,\min} = \min_{s,a} N_k^{\vartheta}(s,a)$. For any $\{\epsilon_k\}_{k=1}^K$, with probability at least $1-\delta$, we take*

$$\hat{c}_{k,h} = \sum_{\vartheta \in \{\alpha,\beta\}} H\sqrt{\frac{2S\ln(16SAHk^2/(\epsilon_k\delta))}{N_k^{\vartheta,\min}}} + 2\epsilon_k HS.$$

Then, the sequence $\{\hat{c}_{k,h}\}_{k,h}$ is compatible.

Similarly, for $\hat{d}_{k,h}$, we have the following lemma.

Lemma 4.3. *Denote $p_k^{\vartheta,\min} = \min_{s,a} p_k^{\vartheta}(y=1|s,a)$. For any $\{\epsilon_k\}_{k=1}^K$, with probability at least $1-\delta$, we take*

$$\hat{d}_{k,h} = \sum_{\vartheta \in \{\alpha,\beta\}} \frac{3H\sqrt{\frac{2S\ln(32SAk^2/(\epsilon_k\delta))}{N_k^{\vartheta,\min}}} + 3\epsilon_k HS}{p_k^{\vartheta,\min}\left(p_k^{\vartheta,\min} - \sqrt{\frac{4\ln 2 + 2\ln(4SAk^2/\delta)}{N_k^{\vartheta,\min}}}\right)}$$

if $p_k^{\vartheta,\min} > \sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,\min}}}$; Otherwise, we set $\hat{d}_{k,h} = 1$. Then, the sequence $\{\hat{d}_{k,h}\}_{k,h}$ is compatible.

4.2 Main Theorems

In this subsection, we provide our formal theoretical guarantees for the reward regret and fairness constraints violation. We require technical Assumption 4.1.

Assumption 4.1. *(a). For all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, there exists a universal constant $C > 0$, such that $p^{*\vartheta}(s'|s,a) \geq C$ for $\vartheta = \{\alpha,\beta\}$. (b). For all $(x,a) \in \mathcal{X}\times\mathcal{A}$, there exists a universal constant $\tilde{C}$, such that $\pi^{*\vartheta}(a|x) \geq \tilde{C}$.*

Assumption 4.1 implies irreducibility of the Markov process and ensures feasibility of our optimization problems (see Appendix C). Recall that at episode $k \in [K]$, for each group ϑ, we have $n_k^{\vartheta} \geq 1$ individuals drawn for each group ϑ. And as a concrete exemplified choice of η_k and ϵ_k, we take $\eta_k = k^{-\frac{1}{3}}$ and $\epsilon_k = (kHS)^{-1}$.

Reward regret. For the reward regret, for either demographic parity or equalized opportunity, we can provide the following theoretical guarantee. Recall for two positive sequences $\{a_j\}$ and $\{b_j\}$, we write $a_j = \mathcal{O}(b_j)$ if $\lim_{j\to\infty}(a_j/b_j) < \infty$.

Theorem 4.1. *For type $\in \{DP, EqOpt\}$, with probability at least $1-\delta$, there exists a threshold $T = \mathcal{O}\left(\left(\frac{H\ln(SA/\delta)}{n_k}\right)^3\right)$, such that for all $k \geq T$,*

$$\mathcal{R}_{\text{reg}}^{type}(k) = \mathcal{O}\left(Hk^{-\frac{1}{3}}\sqrt{HS\ln(S^2AH^2k^3/\delta)}\right).$$

By Theorem 4.1, our algorithms for each of the group fairness notions can ensure vanishing reward regrets when k goes to infinity, which implies that the performance in regard to regret reward improves as the number of episodes increases.

Fairness constraints violation. For fairness violation, we have the following Theorem 4.2.

Theorem 4.2. *For type $\in \{DP, EqOpt\}$, with probability at least $1-\delta$, there exists a threshold $T = \mathcal{O}\left(\left(\frac{H\ln(SA/\delta)}{n_k}\right)^3\right)$, such that for all $k \geq T$,*

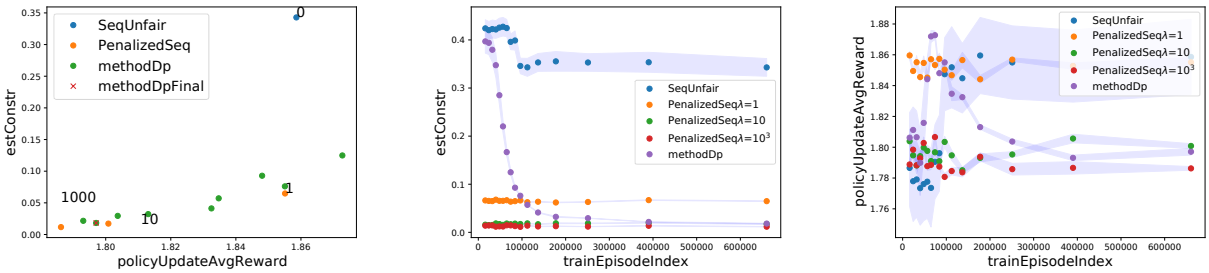
$$\mathcal{C}_{\text{reg}}^{type}(k) \leq \mathcal{O}\left(k^{-\frac{1}{3}}\sqrt{SH\ln(S^2HAk^3/\delta)}\right).$$

By Theorem 4.2, our algorithms for each of the group fairness notions can ensure vanishing fairness violation when k goes to infinity, which implies the performance in regard to fairness violations improves as the number of episodes increases.

5 EXPERIMENTS

Settings. We take $H = 8$ for each episode and update our policy every $k = 2^l$ episodes, where $l = 3, 4, \ldots, 18$. This update schedule helps to reduce computational cost by reducing the number of optimization problems we need to solve while still collecting a large quantity of data. We choose the relaxation parameters $\hat{c}_{k,h}$ and $\hat{d}_{k,h}$ as defined in the previous sections. After each policy update, we use 8,000 episodes to evaluate the new policy. All confidence intervals come from repeating each experiment five times.

Estimation and Optimization Process. We estimate transition probabilities and rewards using the counting method outlined above. These estimates are used as the input for our algorithm. The optimization problems are non-convex and the detailed optimization formulations are included in Appendix F. We use the Gurobi optimization package (Gurobi Optimization, LLC, 2023) to solve, and set the optimality-value tolerance, feasibility



(a) DP FICO Pareto (b) DP FICO estConstr (c) DP FICO return

(d) EqOpt FICO Pareto (e) EqOpt FICO estConstr (f) EqOpt FICO return

Figure 1: FICO data result (methodDp and methodEqOpt stand for our proposed methods). 1a and 1d give the Pareto frontier, where the x axis is the average episodic return and the y axis is the constraint violation level. The point with the cross marker is the result for the final episode and the text annotation provide the penalty parameters. 1b and 1c give the constraint violation level during training. 1e and 1f give the average episodic during training.

tolerance, and solving time limit as 1e-3, 1e-6, and 300 seconds, respectively. We use the barrier algorithm for all problems.

FICO Data. We adapt FICO score data[3] to form our data generating process. The FICO score dataset contains data from non-Hispanic White and Black cohorts consist of $j \in \mathcal{X} = \{0, \ldots, 4\}$, which represent normalized scores $\{0, 25, 50, 75, 100\}$ and can be viewed as the feature. In particular, the FICO data provides empirical distributions for credit scores of different sensitive groups, i.e. $\hat{\mathbb{P}}_{\text{FICO}}(x^\vartheta = j)$, along with an empirical qualification distribution conditioned on each score level $\hat{\mathbb{P}}_{\text{FICO}}(y^\vartheta = y | x^\vartheta = j)$, where $y \in \{0, 1\}$. We simulate the data generating process according to the empirical distributions stated above. For the population dynamics, we model the intial score distribution as $\mathbb{P}(x_0^\vartheta = j) = \hat{\mathbb{P}}_{\text{FICO}}(x_0^\vartheta = j)$, according to the empirical FICO distribution. For the initial qualification distribution conditioned on score, $\mathbb{P}(y_0^\vartheta = 1 | x_0 = j) = \hat{\mathbb{P}}_{\text{FICO}}(y^\vartheta = 1 | x^\vartheta = j)$. Then, we generate the underlying group-dependent and time invariant transition kernel $p^{*\vartheta}$ in the following way: we first set a distribution for $p^{*\vartheta}(x' = j' | x = j, y = w, a = v)$ for $j', j \in \mathcal{X}$ and $w, v \in \{0, 1\}$. Then, we set $p^{*\vartheta}(x', y' | x, y, a)$ as $p^{*\vartheta}(x' | x, y, a) \hat{\mathbb{P}}_{\text{FICO}}(y^\vartheta = y' | x^\vartheta = x')$. More details about the data generating process are given in Appendix F.

[3] https://docs.responsibly.ai/notebooks/demo-fico-analysis.html

Reward function. We choose a score-dependent reward, conditioned on the qualification level and decision as the following:

$$r^\vartheta(x_h, y_h, a_h) = \begin{cases} \beta_1^\vartheta x_h, & \text{if } y_h = 1, a_h = 1; \\ -\beta_2^\vartheta x_h, & \text{if } y_h = 0, a_h = 1; \\ 0, & \text{if } a_h = 0, \end{cases}$$

where $\beta_1^\vartheta, \beta_2^\vartheta \in (0, 1)$. This reward function captures the idea that a decision maker can give a higher loan amount to a candidate with a higher credit score. As a result, the decision maker benefits more when a more qualified candidate has a higher score; i.e., gaining a larger total repayment interest from a qualified candidate with higher score because of making a higher loan amount. The decision maker also suffers a larger, negative reward if the candidate with higher score is not qualified, because a larger loan goes unpaid. In our experiment, we set $\beta_1^\alpha = 0.1, \beta_1^\beta = 0.9, \beta_2^\alpha = 0.9, \beta_2^\beta = 0.1$. This

difference in reward function per group further reflects a decision maker who may, potentially irrationally and unfairly, benefit or penalize through their reward function for different groups, even for two individuals who otherwise have the same score x and qualification y (perhaps the decision maker worries that the model is not equally accurate or calibrated per group).

Baselines. We use the following policies as baselines:

1. SeqUnfair: Sequentially optimal policies without fairness constraints ($\lambda = 0$).

2. PenalizedSeq: Sequentially optimal policies with an objective that includes a fairness penalty term, which serves as an alternative method for our proposed optimization problem. In this case, our optimization objective, for a particular kind of fairness constraint, is:

$$\max_{\boldsymbol{\pi}\in\Pi_k} \sum_{h=1}^{H} \left[\mathcal{R}_{k,h}(\boldsymbol{p}_k, \boldsymbol{\pi}) + FP(\{\pi^\vartheta, p_k^\vartheta, s_h^\vartheta, a_h^\vartheta\}_\vartheta; \lambda)\right],$$

where $FP(\{\pi^\vartheta, p_k^\vartheta, s_h^\vartheta, a_h^\vartheta\}_\vartheta; \lambda)$ is the fairness penalty and λ is the penalty parameter ($\lambda = 1, 10, 10^3$).

(a). *Demographic parity*. For this case, we choose $FP(\{\pi^\vartheta, p_k^\vartheta, s_h^\vartheta, a_h^\vartheta\}_\vartheta; \lambda)$ as

$$\lambda(\mathbb{P}^{\pi^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi^\beta, p_k^\beta}(a_{k,h}^\beta = 1))^2.$$

(b). *Equalized opportunity.* For this case, we choose $FP(\{\pi^\vartheta, p_k^\vartheta, s_h^\vartheta, a_h^\vartheta\}_\vartheta; \lambda)$ as

$$\lambda\Big(\mathbb{P}^{\pi^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1 | y_{k,h}^\alpha = 1) - \mathbb{P}^{\pi^\beta, p_k^\beta}(a_{k,h}^\beta = 1 | y_{k,h}^\beta = 1)\Big)^2.$$

Experimental results. Figure 1a shows the Pareto frontier in terms of episodic total return and episodic step-average fairness violation for demographic parity, and Figure 1d shows the counterpart for equal opportunity. Figure 1b and 1c demonstrate the training dynamics of different algorithms for demographic parity, and Figures 1e and 1f demonstrate the counterpart for equal opportunity. Our proposed method converges to a stable level in terms of fairness violation over the training episodes. In addition, from the confidence intervals (the shaded area in the graph), we can see that our algorithm has a much narrower confidence band than the other baseline.

6 DISCUSSION

In this section, we discuss possible extensions of our framework and future directions.

6.1 Extension of Stepwise Fairness Notions

We focus on two popular common types of fairness criteria, namely demographic parity and equalized opportunity. However, our techniques can also be extended in future work to additional types of fairness criteria. In particular, we can consider a family of fairness constraints, as formally introduced in Agarwal et al. (2018). They consider constraints in the form

$$M\mu(a) \leq c$$

for matrix M and vector c, where the j-th coordinate of μ is $\mu_j(a) = \mathbb{E}[g(x, y, a, \vartheta)|E_j]$ for $j \in \mathcal{J}$, and $M \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{J}|}$, $c \in \mathbb{R}^{\mathcal{K}}$. Here, $\mathcal{K} = \mathcal{A} \times \mathcal{Y} \times \{+, -\}$ ($+, -$ impose positive/negative sign so as to recover $|\cdot|$ in constraints), for $\mathcal{Y} = \{0, 1\}$, and $\mathcal{J} = (\Lambda \cup \{*\}) \times \{0, 1\}$. E_j is an event defined with respect to (x, y, ϑ) and $*$ denotes the entire probability space. This formulation includes demographic parity and equalized opportunity as special cases.

6.2 Aggregated Fairness Notions

Our techniques can also be extended in future work to adopt aggregate fairness notions that consider the entire episodic process. Specifically, we could consider *aggregate equalized opportunity* (also called *aggregate true positive rate* in D'Amour et al. (2020))

$$\sum_{h=1}^{H} \mathbb{P}(a_h = 1 | y_h = 1, \vartheta) \frac{\mathbb{P}(y_h = 1|\vartheta)}{\sum_{h=1}^{H} \mathbb{P}(y_h = 1|\vartheta)},$$

which can be viewed as a weighted sum of equalized opportunity across steps. This should be relatively straightforward to handle with our techniques.

6.3 Non-episodic, Infinite Horizon Markov Decision Processes

Another natural direction is to extend our framework to non-episodic infinite horizon. Taking DP as an example, we could consider

$$\max_{\boldsymbol{\pi}\in\Pi_k} \sum_{h=1}^{\infty} \gamma^h \mathcal{R}_h(\boldsymbol{p}, \boldsymbol{\pi}),$$
$$s.t.\ \forall h \in [H],$$
$$\gamma^h |\mathbb{P}^{\pi^\alpha, p^\alpha}(a_h^\alpha = 1) - \mathbb{P}^{\pi^\beta, p^\beta}(a_h^\beta = 1)| \leq \hat{c}_h.$$

This will involve using a more advanced version of concentration inequalities for Markov chains, and we leave this to future work.

7 CONCLUSION

In this paper, we have introduced the study of reinforcement learning with stepwise fairness constrains, which are defined to require group fairness criteria to be satisfied at each time step. We have provided learning algorithms with theoretical guarantees in regard to policy optimality and fairness violations for the case of tabular episodic RL. Our claims are well-supported by the experimental results.

ACKNOWLEDGEMENTS

The research of Linjun Zhang is partially supported by NSF DMS-2015378. Zhiwei Steven Wu is supported in part by the NSF FAI Award 1939606.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Ian Ball. Scoring strategic agents. In *Arxiv*, 2019.

Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*, 2020.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, ICDMW'09*, 2009.

Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *Arxiv*, 2019.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. *arXiv preprint arXiv:2206.02792*, 2022.

Zhun Deng, Cynthia Dwork, and Linjun Zhang. Happymap: A generalized multicalibration method. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2023.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer-Reingold, and Richard Zemel. Fairness through awarenes. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, 2018.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1617–1626, 2017.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Leonid Aryeh Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Matt J Kusner, Joshua Loftus, Chris Russell, , and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. In *Arxiv*, 2022.

Yonadav Shavit and William S. Moses. Extracting incentives from black-box decisions. In *Arxiv*, 2019.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469, 2020.

Supplementary Materials

A SOLVING OUR ALGORITHMS VIA OCCUPANCY MEASURES

We use occupation measure to reformulate the problem, so that the objective is stated in terms of occupation measures. For episode k, we denote the $\mathbb{P}^{\pi_k^\vartheta, p_k^\vartheta}(x_{k,h}^\vartheta = x, y_{k,h}^\vartheta = y, a_{k,h}^\vartheta = a)$ as $\rho_k^\vartheta(x, y, a, h)$, which we call occupancy measures.

A.1 Demographic Parity

For episode k, the optimization problem can be reformulated as:

$$\max_{\rho} \sum_{x,y,a,h,\vartheta} p_\vartheta \rho_k^\alpha(x, y, a, h) \hat{r}_k^\vartheta(x, y, a)$$

such that

$$\forall h \quad \left| \sum_{y,x} \rho_k^\alpha(x, y, a = 1, h) - \sum_{y,x} \rho_k^\beta(x, y, a = 1, h) \right| \le \hat{c}_{k,h},$$

$$\forall \vartheta, x, t \quad \frac{\rho_k^\vartheta(x, y = 1, a = 1, h)}{\sum_a \rho_k^\vartheta(x, y = 1, a, h)} = \frac{\rho_k^\vartheta(x, y = 0, a = 1, h)}{\sum_a \rho_k^\vartheta(x, y = 0, a, h)} \quad \text{this formula makes sure the policy only depends on } x$$

$$\forall \vartheta, x', y', h \quad \sum_a \rho_k^\vartheta(x', y', a, h + 1) = \sum_{x,y,a} \rho_k^\vartheta(x, y, a, h) p_{(k)}^\vartheta(x', y' | x, y, a),$$

$$\forall \vartheta, x, y, a, h \;\; 0 \le \rho_k^\vartheta(x, y, a, h) \le 1, \quad \sum_{x,y,a} \rho_k^\vartheta(x, y, a, h) = 1.$$

A.2 Equal Opportunity

For episode k, the optimization problem can be reformulated as:

$$\max_{\rho} \sum_{x,y,a,h,\vartheta} p_\vartheta \rho_k^\alpha(x, y, a, h) \hat{r}_k^\vartheta(x, y, a)$$

such that

$$\forall h \quad \left| \frac{\sum_x \rho_k^\alpha(x, y = 1, a = 1, h)}{\sum_{x,a} \rho_k^\alpha(x, y = 1, a, h)} - \frac{\sum_x \rho_k^\beta(x, y = 1, a = 1, h)}{\sum_{x,a} \rho_k^\beta(x, y = 1, a, h)} \right| \le \hat{d}_{k,h},$$

$$\forall \vartheta, x, t \quad \frac{\rho_k^\vartheta(x, y = 1, a = 1, h)}{\sum_a \rho_k^\vartheta(x, y = 1, a, h)} = \frac{\rho_k^\vartheta(x, y = 0, a = 1, h)}{\sum_a \rho_k^\vartheta(x, y = 0, a, h)} \quad \text{this formula makes sure the policy only depends on } x$$

$$\forall \vartheta, x', y', h \quad \sum_a \rho_k^\vartheta(x', y', a, h + 1) = \sum_{x,y,a} \rho_k^\vartheta(x, y, a, h) p_{(k)}^\vartheta(x', y' | x, y, a),$$

$$\forall \vartheta, x, y, a, h \;\; 0 \le \rho_k^\vartheta(x, y, a, h) \le 1, \quad \sum_{x,y,a} \rho_k^\vartheta(x, y, a, h) = 1.$$

B EXTENSION TO MULTIPLE SENSITIVE ATTRIBUTES

Consider multiple attributes $\vartheta \in \Theta = \{\theta_1, \theta_2, \cdots, \theta_q\}$, where Θ is a set of sensitive attributes. We denote the expected reward for a random individual at time step h as,

$$\mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi}) = \sum_{\vartheta \in \Theta} p_\vartheta \cdot \mathbb{E}^{\pi^\vartheta, p^{*\vartheta}}[r^{*\vartheta}(s_h^\vartheta, a_h^\vartheta)].$$

Here, $\mathbb{E}^{\pi^\vartheta, p^{*\vartheta}}[\cdot]$ refers to the expected value for an individual in group ϑ, i.e., conditioned on the individual being sampled in this group. Our goal is to obtain the policy $\boldsymbol{\pi}^*$ that solves the following optimization problem:

$$\begin{gathered}\max_{\boldsymbol{\pi}} \sum_{h=1}^{H} \mathcal{R}_h^*(\boldsymbol{p}^*, \boldsymbol{\pi}) \\ \text{s.t. } \forall h \in [H], \\ faircon(\{\pi^\vartheta, p^{*\vartheta}, s_h^\vartheta, a_h^\vartheta\}_{\vartheta \in \Theta}),\end{gathered}$$

where $faircon$ corresponds to a particular fairness concept.

(i) *RL with demographic parity (DP).* For this case, $faircon$ is for any $\theta_i, \theta_j \in \Theta$

$$\mathbb{P}^{\pi^{\theta_i}, p^{*\theta_i}}[a_h^{\theta_j} = 1] = \mathbb{P}^{\pi^\beta, p^{*\theta_j}}[a_h^{\theta_j} = 1],$$

which means at each time step h, the decision for individuals from different groups is statistically independent of the sensitive attribute.

(ii) *RL with equalized opportunity (EqOpt).* For this case, $faircon$ is for any $\theta_i, \theta_j \in \Theta$

$$\mathbb{P}^{\pi^{\theta_i}, p^{*\theta_i}}[a_h^{\theta_j} = 1 | y_h^{\theta_j} = 1] = \mathbb{P}^{\pi^\beta, p^{*\theta_j}}[a_h^{\theta_j} = 1 | y_h^{\theta_j} = 1],$$

which means at each time step h, the decision for a random individual from each of the two different groups is statistically independent of the sensitive attribute conditioned on the individual being qualified.

The corresponding practical optimization for DP and EqOpt can also adapted to multiple sensitive attributes similarly.

C FEASIBILITY DISCUSSIONS

Let us recall the following assumption.

Assumption C.1 (Restatement of Assumption 4.1)**.** *(a). For all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, there exists a universal constant $C > 0$, such that $p^{*\vartheta}(s'|s, a) \geq C$ for $\vartheta = \{\alpha, \beta\}$. (b). For all $(x, a) \in \mathcal{X} \times \mathcal{A}$, there exists a universal constant $\tilde{C}$, such that $\pi^{*\vartheta}(a|x) \geq \tilde{C}$.*

For the ideal optimization, Assumption 4.1 makes sure that every step we have $\mathbb{P}^{\pi^\vartheta, p^{*\vartheta}}(s_h^\vartheta = s, a_h^\vartheta = a) \geq C \min_x \pi^\vartheta(a|x)$ for all s, a and $h > 1$. Thus, for DP, we can set $\pi^\alpha(a = 1|x) = \pi^\beta(a = 1|x) = 1$ for all x, then, we know this policy is a feasible policy. Similarly for EqOpt, we can also take $\pi^\alpha(a = 1|x) = \pi^\beta(a = 1|x) = 1$ for all x, and this is a feasible solution given the event $y_h^\vartheta = 1$ is always of positive probability.

The same argument can be applied to the practical optimization for both DP and EqOpt.

D MODELLING INDIVIDUAL'S OPTING IN/OUT

Recall that in order to analyze the policy's effect on the population, we model the behavior of a randomly drawn individual who interacts with the environment across H steps. However, when we gather data from real individuals, they may opt in opt out at a certain step and not interact with the environment throughout the H steps. Nevertheless, we can still aggregate their data and provide estimation of quantities of interest – a representative randomly drawn individual that interact with the environment for the full episode.

Specifically, at episode $k \in [K]$, each time step $h \in [H]$, there are $n_{k,h}$ individuals (people opt in and opt out in the process, so there are different number of people at each time step). For each group ϑ, we pool all those people together and obtain $n_k^\vartheta = \sum_h n_{k,h}^\vartheta$ people in total. We will use the counting method to obtain empirical measures for those quantities, each quantity will need to sum over all the n_k^ϑ people. For the i-th individual, his/her status and action at time h is denoted as $s_{k,h}^{\vartheta,i}$ and $a_{k,h}^{\vartheta,i}$. Some people will opt out, for example the i-th individual opts out at $h+1$, so there is no status for him/her at time $h+1$. Given that, let us define $\mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a, s_{k,h+1}^{\vartheta,i} = \cdot\}$ to be the indicator, which will be 1 only when $s_{k,h}^{(i)} = s, a_{k,h}^{(i)} = a$, and that individual still hasn't opted out at time $h+1$ ($s_{k,h+1}^{\vartheta,i}$ exists). Similarly, if the the i-th individual has opted out at time $h+1$, those indicators used below concerning $s_{k,h+1}^{\vartheta,i}$ will be 0.

For $\vartheta \in \{\alpha, \beta\}$, let $\boldsymbol{p}_k = (p_k^\alpha, p_k^\beta)$ and $\boldsymbol{r}_k = (r_k^\alpha, r_k^\beta)$,

$$N_k^\vartheta(s,a) = \max\big\{1, \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} \mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a, s_{k,h+1}^{\vartheta,i} = \cdot\}\big\},$$

$$p_k^\vartheta(s'|s,a) = \frac{1}{N_k^\vartheta(s,a)} \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} \mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a, s_{k,h+1}^{\vartheta,i} = s'\},$$

$$\hat{r}_k^\vartheta(s,a) = \frac{1}{N_k^\vartheta(s,a)} \sum_{t\in[k-1],h\in[H],i\in[n_k^\vartheta]} r^\vartheta(s,a)\mathbf{1}\{s_{k,h}^{\vartheta,i} = s, a_{k,h}^{\vartheta,i} = a\}.$$

Our analysis carries over to this setting as long as we assume at each time step h in all episodes there is an individual who will not opt out in the next time step $h+1$.

E OMITTED PROOFS

For simplicity of notation, we **omit the superscript** ϑ in most of the proofs. In addition, without loss of generality, **we assume** $r \in [0,1]$. This is just for proof simplicity, and our algorithms can still be applied to settings with negative reward values. Also, for quantities defined below such as $Q_r^{\pi,p}$ and $V_r^{\pi,p}$, we will **omit subscripts and superscripts** when it is clear from the text.

Recall the following concepts:

Q and V functions. Two of the most common concepts in RL are Q and V functions. Specifically, Q functions track the expected reward when a learner starts from state $s \in \mathcal{S}$. Meanwhile, V functions are the corresponding expected Q functions of the selected action. For a reward function r and a MDP with transition function p, Q and V functions are defined as:

$$Q_r^{\pi,p}(s,a,h) = r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)V_r^{\pi,p}(s',h+1),$$
$$V_r^{\pi,p}(s,h) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_r^{\pi,p}(s,a,h)],$$

where we set $V_r^{\pi,p}(s,H+1) = 0$.

Bellman error. For an arbitrary function m, for underlying objectives m^* and p^*, the *Bellman error* for p, m under policy π at stage h is denoted as:

$$\mathcal{B}_m^{\pi,p}(s,a,h) = Q_m^{\pi,p}(s,a,h) - \Big(m^*(s,a) + \sum_{s'} p^*(s'|s,a)V_m^{\pi,p}(s',h+1)\Big). \tag{1}$$

E.1 Proof of lemma 4.1

Lemma E.1 (Restatement of lemma 4.1)**.** *If we take* $\hat{b}_k^\vartheta = \min\Big\{2H, 2H\sqrt{\frac{2\ln(16SAHk^2/\delta)}{N_k^\vartheta(s,a)}}\Big\}$*, then with probability* $1-\delta$*, the bonus* $\hat{b}_k^\vartheta(s,a)$ *is valid for all* k *episodes and* $\vartheta = \{\alpha, \beta\}$.

Proof. For a fixed $\vartheta \in \{\alpha, \beta\}$, given that $r \in [0,1]$, we have $\sup_{s\in\mathcal{S},h\in H} |V(s,h)| \le H$. For a single state-action pair (s,a), by Hoeffding inequality, with probability $1-\delta'$

$$\left|\hat{r}_k(s,a) - r_k^*(s,a) + \sum_{s'\in\mathcal{S}} \big(p_k(s'|s,a) - p^*(s'|s,a)\big)V(s',h+1)\right| \le 2H\sqrt{\frac{2\ln(2/\delta')}{N_k(s,a)}}.$$

Also, by the boundedness of $\hat{r}_k$ and V,

$$\left|\hat{r}_k(s,a) - r_k^*(s,a) + \sum_{s'\in\mathcal{S}} \big(p_k(s'|s,a) - p^*(s'|s,a)\big)V(s',h+1)\right| \le 2H.$$

Further if we take $\delta' = \frac{\delta}{4SAHk^2}$, and further apply union bound on stats and actions, then the failure probability is $\delta/(4k^2)$ for episode k. Lastly, bounding across episodes, we have the failure probabilty is:

$$\sum_{i=1}^{K} \frac{\delta}{4k^2} \le \delta.$$

Since there are two values for ϑ, again we apply union bound, then we have the final result. □

E.2 Proof of Lemma 4.2

To prove our result, we need the following lemma.

Lemma E.2 (Simulation lemma (Kearns and Singh, 2002)). *For any policy π, objective m, transition probabilty p, and underlying objectives m^*, p^*, it holds that*

$$\mathbb{E}^\pi V_m^{\pi,p}(s_1,1) - \mathbb{E}^\pi V_{m^*}^{\pi,p^*}(s_1,1) = \mathbb{E}^\pi\Big[\sum_{h=1}^{H} \mathcal{B}_m^{\pi,p}(s_h,a_h,h)\Big].$$

Lemma E.3 (Restatement of Lemma 4.2). *Denote $N_k^{\vartheta,min} = \min_{s,a} N_k^\vartheta(s,a)$, for any $\{\epsilon_k\}_{k=1}^K$, with probability at least $1-\delta$, we take*

$$\hat{c}_{k,h} = \sum_{\vartheta\in\{\alpha,\beta\}} H\sqrt{\frac{2S\ln(16SAHk^2/(\epsilon_k\delta))}{N_k^{\vartheta,min}}} + 2\epsilon_k HS,$$

then $\hat{c}_{k,h}$ is compatible for all $h \in [H]$ and $k \in [K]$.

Proof. For a specific time step h^*, let us consider taking

$$m_{h^*}(s_{k,h},a_{k,h}) = m_{h^*}^*(s_{k,h},a_{k,h}) = \begin{cases} \mathbf{1}\{a_{k,h}=1\}, & \text{if } h = h^*. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

And it is easy to observe that

$$\mathbb{E}^\pi V_{m_{h^*}}^{\pi,p}(s_1,1) = \mathbb{P}^{\pi,p}(a_{k,h^*}=1).$$

Thus, in order to bound

$$|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1) - \mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1)|,$$

it is equivalent is to bound

$$|\mathbb{E}^{\pi^*} V_{m_{h^*}}^{\pi^*,p_k}(s_1,1) - \mathbb{E}^{\pi^*} V_{m_{h^*}}^{\pi^*,p^*}(s_1,1)| = \left|\mathbb{E}^{\pi^*}\Big[\sum_{h=1}^{H} \mathcal{B}_{m_{h^*}}^{\pi^*,p_k}(s_h,a_h,h)\Big]\right|.$$

Here, a slight fine-grained analysis suggests that we can replace $\sum_{h=1}^{H}$ to $\sum_{h=1}^{h^*}$, but this change cannot change the order of the final bound, so for simplicity, we still use $\sum_{h=1}^{H}$.

Now, let us analyze $\left|\mathcal{B}^{\pi^*,p_k}_{m_{h^*}}(s_h,a_h,h)\right|$. Specifically,

$$\left|\mathcal{B}^{\pi^*,p_k}_{m_{h^*}}(s_h,a_h,h)\right| = \Big|\sum_{s'\in\mathcal{S}}\Big(p_k(s'|s,a)-p^*(s'|s,a)\Big)V^{\pi^*,p_k}_{m_{h^*}}(s')\Big|.$$

Since $V^{\pi^*,p_k}_{m_{h^*}}(s')$ (we will use V for simplicity from now on) is data dependent, we need to to use a union bound argument. Notice $V(s)\in[0,1]$ for all s, thus, we can let Ψ to be a ϵ-net on $[0,1]^S$. For any fixed $V\in\Psi$, by similar proof as in Lemma 4.1, we have with probability at least $1-\delta'$, for all $k\in[K]$,

$$\Big|\sum_{s'\in\mathcal{S}}\Big(p_k(s'|s,a)-p^*(s'|s,a)\Big)V(s')\Big| \le \sqrt{\frac{2\ln(8SAk^2/\delta')}{N_k(s,a)}}.$$

The cardinality of Ψ is $(1/\epsilon)^S$. Thus, by using union bound and let $\delta=\delta'/(1/\epsilon)^S$ (we don't need to union bound over H because we have bounded for all elements in epsilon nets), we have with probability at least $1-\delta$, for all $h\in[H]$,

$$\left|\mathcal{B}^{\pi^*,p_k}_{m_{h^*}}(s_h,a_h,h)\right| \le \sqrt{\frac{2S\ln(8SAHk^2/(\epsilon\delta))}{N_k(s_h,a_h)}}+\epsilon S.$$

Notice our argument still valid if we set ϵ as ϵ_k for episode k. Then, we have that with probability at least $1-\delta$,

$$\hat{c}_{k,h}=\sum_{\vartheta\in\{\alpha,\beta\}}2\sqrt{\frac{2S\ln(16SAk^2/(\epsilon_k\delta))}{\min_{s,a}N_k^{\vartheta}(s,a)}}+2\epsilon_k HS.$$

is compatible for all $k\in[K]$.

□

E.3 Proof of Lemma 4.3

Lemma E.4 (Restatement of Lemma 4.3)**.** *Denote $p_k^{\vartheta,min}=\min_{s,a}p_k(y=1|s,a)$, For any $\{\epsilon_k\}_{k=1}^K$, with probability at least $1-\delta$, we take*

$$\hat{d}_{k,h}=\sum_{\vartheta\in\{\alpha,\beta\}}\frac{3H\sqrt{\frac{2S\ln(32SAk^2/(\epsilon_k\delta))}{N_k^{\vartheta,min}}}+3\epsilon_k HS}{p_k^{\vartheta,min}\left(p_k^{\vartheta,min}-\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}}\right)}$$

if $p_k^{\vartheta,min}>\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}}$; Otherwise, we set $\hat{d}_{k,h}=1$. Then $\hat{d}_{k,h}$ is compatible for all $h\in[H]$ and $k\in[K]$.

Proof. By Bretagnolle-Huber-Carol's inequality, we know that with probability at least $1-\delta$, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $k\in[K]$

$$\sum_{y\in\mathcal{Y}}|p_k(y|s,a)-p^*(y|s,a)|\le\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{N_k(s,a)}}.$$

By proof as in Lemma 4.2, we know that by taking

$$m_{h^*}(s_{k,h},a_{k,h})=m^*_{h^*}(s_{k,h},a_{k,h})=\begin{cases}\mathbf{1}\{a_{k,h}=1,y_{k,h}=1\}, & \text{if } h=h^*,\\ 0, & \text{otherwise,}\end{cases} \tag{3}$$

we have for any $\epsilon>0$, for any $a\in\mathcal{A}$

$$|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=a,y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=a,y_{k,h^*}=1)|\le H\sqrt{\frac{2S\ln(8SAk^2/(\epsilon\delta))}{\min_{s,a}N_k(s,a)}}+\epsilon HS,$$

with probability at least $1-\delta$.

On the other hand, by summing over a for $\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=a,y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=a,y_{k,h^*}=1)$, we have

$$\begin{aligned}|\mathbb{P}^{\pi^*,p_k}(y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(y_{k,h^*}=1)| &\le \sum_a |\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=a,y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=a,y_{k,h^*}=1)|\\ &\le 2H\sqrt{\frac{2S\ln(8SAk^2/(\epsilon\delta))}{\min_{s,a}N_k(s,a)}}+2\epsilon HS,\end{aligned}$$

If $h>1$

$$\begin{aligned}\mathbb{P}^{\pi^*,p^*}(y_{k,h}=1) &\ge \sum_{s,a} p^*(y_{k,h}=1|s_{k,h-1}=s,a_{k,h-1}=a)\mathbb{P}^{\pi^*,p^*}(s_{k,h-1}=s,a_{k,h-1}=a)\\ &\ge \min_{s,a} p^*(y=1|s,a).\end{aligned}$$

Since s_1 can be chosen by us, so we can make sure $\mathbb{P}^{\pi^*,p^*}(y_{k,1}=1)$ bounded away from 0. Actually, even we set $y_{k,1}=0$, conditioning on $\emptyset$ automatically satisfy EqOpt constraint.

Similarly,

$$\begin{aligned}\mathbb{P}^{\pi^*,p_k}(y_{k,h}=1) &\ge \sum_{s,a} p_k(y_{k,h}=1|s_{k,h-1}=s,a_{k,h-1}=a)\mathbb{P}^{\pi^*,p_k}(s_{k,h-1}=s,a_{k,h-1}=a)\\ &\ge \min_{s,a} p_k(y=1|s,a).\end{aligned}$$

By Bretagnolle-Huber-Carol's inequality, with probability at least $1-\delta$, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$

$$|p_k(y=1|s,a)-p^*(y=1|s,a)|\le\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{N_k(s,a)}}.$$

As a result, if

$$\min_{s,a} p_k(y=1|s,a)>2\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{\min_{s,a}N_k(s,a)}},$$

then, with probability at least $1-\delta$,

$$\min_{s,a} p^*(y=1|s,a)>\min_{s,a} p_k(y=1|s,a)-\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{\min_{s,a}N_k(s,a)}},$$

which further leads to

$$\begin{aligned}&|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1|y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1|y_{k,h^*}=1)|\\ &\le \frac{|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1,y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1,y_{k,h^*}=1)|+|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1)|}{p_k(y=1|s,a)p^*(y=1|s,a)}\\ &\le \frac{|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1,y_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1,y_{k,h^*}=1)|+|\mathbb{P}^{\pi^*,p_k}(a_{k,h^*}=1)-\mathbb{P}^{\pi^*,p^*}(a_{k,h^*}=1)|}{\min_{s,a}p_k(y=1|s,a)(\min_{s,a}p_k(y=1|s,a)-\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{\min_{s,a}N_k(s,a)}})}\\ &\le \frac{3H\sqrt{\frac{2S\ln(8SAk^2/(\epsilon\delta))}{\min_{s,a}N_k(s,a)}}+3\epsilon HS}{\min_{s,a}p_k(y=1|s,a)(\min_{s,a}p_k(y=1|s,a)-\sqrt{\frac{4\ln 2+2\ln(SAk^2/\delta)}{\min_{s,a}N_k(s,a)}})}.\end{aligned}$$

Notice our argument still valid if we set ϵ_k for episode k and apply union bounds for all the events mentioned above and $\vartheta=\{\alpha,\beta\}$, then, we have that with probability at least $1-\delta$

$$\hat{d}_{k,h} = \begin{cases} \sum_{\vartheta} \frac{3H\sqrt{\frac{2S\ln(32SAk^2/(\epsilon_k\delta))}{N_k^{\vartheta,min}}}+3\epsilon_k HS}{p_k^{\vartheta,min}\left(p_k^{\vartheta,min}-\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}}\right)}, & \text{if } p_k^{\vartheta,min} > \sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}}; \\ 1, & \text{otherwise;} \end{cases}$$

is compatible for all $k \in [K]$. □

E.4 Proof of Theorem 4.1

We first need a lemma regarding the lower bound of $\min_{s,a} N_k^{\vartheta}(s,a)$.

Restatement of Result in Kontorovich and Ramanan (2008) We consider a simplified variant of Theorem 1.1 in Kontorovich and Ramanan (2008). Let $Z_i \in S$, where S is a finite set, and $Z = (Z_1, Z_2, \cdots, Z_n)$. We further denote $Z_i^j = (Z_i, Z_{i+1}, \cdots, Z_j)$ as a random vector for $1 \leqslant i < j \leqslant n$. Correspondingly, we let $z_i^j = (z_i, z_{i+1}, \cdots, z_j)$ be a subsequence for $(z_1, z_2, \cdots, z_n)$. And let

$$\bar{\eta}_{i,j} = \sup_{v_1^{i-1} \in S^{i-1}, w, w' \in S,\ \mathbb{P}(Z_1^i = Y^{i-1}w) > 0,\ \mathbb{P}(Z_1^i = V^{i-1}w') > 0} \eta_{i,j}(v_1^{i-1}, w, w'),$$

where

$$\eta_{i,j}(v_1^{i-1}, w, w') = TV\Big(\mathcal{D}(Z_j^n | Z_1^i = v_1^{i-1}w), \mathcal{D}(Z_j^n | Z_1^i = v_1^{i-1}w')\Big).$$

Here TV is the total variational distance, and $\mathcal{D}(Z_j^n | Z_1^i = v_1^i w)$ is the conditional distribution of Z_j^n conditioning on $\{Z_1^i = v_1^i w\}$.

Let H_n be $n \times n$ upper triangular matrix defined by

$$(H_n)_{ij} = \begin{cases} 1 & i = j \\ \bar{\eta}_{i,j} & i < j \\ 0 & o.w. \end{cases}$$

Then,

$$\|H_n\|_\infty = \max_{1 \leqslant i \leqslant n} J_{n,i},$$

where

$$J_{n,i} = 1 + \bar{\eta}_{i,i+1} + \cdots + \bar{\eta}_{i,n},$$

and $J_{n,n} = 1$.

Theorem E.1 (Variant of Result in Kontorovich and Ramanan (2008))**.** *Let f be a L_f-Lipschitz function (with respect to the Hamming distance) on S^n for some constant $L_f > 0$. Then, for any $t > 0$,*

$$\mathbb{P}(|f(Z) - Ef(z)| \geqslant t) \leqslant 2\exp\Big(-\frac{t^2}{2nL_f^2\|H_n\|_\infty^2}\Big).$$

Lemma E.5. *Under Assumption 4.1, with probability at least $1-\delta$, for all $k \in [K]$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\vartheta \in \{\alpha, \beta\}$*

$$N_k^{\vartheta}(s,a) \geq \left(1 - \sqrt{\frac{2H^2\ln(4k^2SA/\delta)}{c\eta_k n_k(H-1)(k-1)}}\right) c\eta_k n_k (H-1)(k-1).$$

Proof. By (a) in Assumption 4.1, we know that for any (s,a), (s',a')

$$\mathbb{P}^{\pi_k^{\vartheta}, p^{*\vartheta}}(s_{k,h}^{\vartheta} = (x,y), a_{k,h}^{\vartheta} = a) \geq \pi_k^{\vartheta}(a_{k,h}^{\vartheta} = a | x_{k,h}^{\vartheta} = x) p^{*\vartheta}(s_{k,h}^{\vartheta} = (x,y) | s_{k,h-1}^{\vartheta} = s', a_{k,h-1}^{\vartheta} = a') \geq C\eta_k.$$

As η_k is decreasing with respect to k, we must have for all $t \leq k$,

$$\mathbb{P}^{\pi_t^{\vartheta}, p^{*\vartheta}}(s_{t,h}^{\vartheta}, a_{t,h}^{\vartheta}) \geq \pi_t^{\vartheta}(a_{t,h}^{\vartheta} | x_{t,h}^{\vartheta}) p^{*\vartheta}(s_{t,h}^{\vartheta} | s_{t,h-1}^{\vartheta}, a_{t,h-1}^{\vartheta}) \geq C\eta_k.$$

Notice each individual is independent thus,

$$\mathbb{E}\frac{N_k(s,a)}{(H-1)(K-1)} \geq C\eta_k n_k.$$

Meanwhile, the data between different episodes are independent, thus, using the notation in Kontorovich and Ramanan (2008), we know $\bar{\eta}_{ij} = 0$ if $|i-j| \geq H$. Notice total variational distance is always bounded by 1, so

$$\|H_n\|_\infty = \max_{1\leqslant i\leqslant n} J_{n,i} \leqslant H.$$

then, by applying Theorrem E.1 about the concentration result for non-stationary Markov chain,

$$\mathbb{P}\left(|N_t(s,a) - \mathbb{E}N_t(s,a)| \geq \varepsilon c n_k \eta_k (H-1)(k-1)\right) \leq 2\exp\left(\frac{-\varepsilon^2 C\eta_k n_k (H-1)(k-1)}{2H^2}\right).$$

In other words, with probability at least $1-\delta'$,

$$N_k(s,a) \geq \left(1 - \sqrt{\frac{2H^2\ln(2/\delta')}{C\eta_k n_k (H-1)(k-1)}}\right) C\eta_k n_k (H-1)(k-1)$$

Taking $\delta' = \delta/(2SAk^2)$, we have with probability at least $1-\delta$, for all $k \in [K]$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, $\vartheta \in \{\alpha, \beta\}$

$$N_k^\vartheta(s,a) \geq \left(1 - \sqrt{\frac{2H^2\ln(4k^2SA/\delta)}{C\eta_k n_k (H-1)(k-1)}}\right) C\eta_k n_k (H-1)(k-1).$$

□

Next, we will use the concept of *optimism* in Brantley et al. (2020) in our proof.

Definition E.1 (Optimism)**.** *We call (p_k, r_k) is optimisitc if*

$$\mathbb{E}\left[V_{r_k}^{\pi^*,p_k}(s_1,1)\right] \geq \mathbb{E}\left[V_{r^*}^{\pi^*,p^*}(s_1,1)\right].$$

Lemma E.6. *If $\hat{b}_k$ is valid in episode k for all k simultaneously, we have*

$$\mathbb{E}\left[V_{r_k}^{\pi^*,p_k}(s_1,1)\right] \geq \mathbb{E}\left[V_{r^*}^{\pi^*,p^*}(s_1,1)\right].$$

Proof. This proof mainly follows Brantley et al. (2020) by using induction.

Since the setting ends at episode H,

$$Q_{r_k}^{\pi^*,p_k}(s,a,H+1) = Q_{r^*}^{\pi^*,p^*}(s,a,H+1) = 0.$$

We assume that the inductive hypothesis $Q_{r_k}^{\pi^*,p_k}(s,a,h) \geq Q_{r^*}^{\pi^*,p^*}(s,a,h+1)$ (thus, $V_{r_k}^{\pi^*,p_k}(s,h+1) \geq V_{r^*}^{\pi^*,p^*}(s,h+1)$ holds)). Then, for h,

$$\begin{aligned} Q_{r_k}^{\pi^*,p_k}(s,a,h+1) &= r_k + \sum_{s'\in\mathcal{S}} p_k(s'|s,a) V_{r_k}^{\pi^*,p_k}(s',h+1) \\ &\geq r_k + \sum_{s'\in\mathcal{S}} p_k(s'|s,a) V_{r^*}^{\pi^*,p^*}(s',h+1). \end{aligned}$$

Meanwhile, we know,

$$Q_{r^*}^{\pi^*,p^*}(s,a,h) = r^*(s,a) + \sum_{s'\in\mathcal{S}} p^*(s'|s,a) V_{r^*}^{\pi^*,p^*}(s',h+1).$$

Subtracting the above two formulas, we have

$$Q^{\pi^*,p_k}_{r_k}(s,a,h) - Q^{\pi^*,p^*}_{r^*}(s,a,h) \geq (\hat{r}_k(s,a) + \hat{b}_k(s,a) - r^*(s,a)) + \sum_{s'\in\mathcal{S}} (p_k(s'|s,a) - p^*(s'|s,a)) V^{\pi^*,p^*}_{r^*}(s',h+1) \geq 0$$

where the last inequality holds because of the bonuses are valid. □

Let us summarize briefly and informally: with probability $1-4\delta$, we have:

1. $\{c_{k,h}\}_{k,h}$ are compatible;
2. $\{d_{k,h}\}_{k,h}$ are compatible;
3. $\{b^{\vartheta}_{k,h}\}_{k,h}$ are valid;
4. $N^{\vartheta}_k(s,a) \geq \left(1 - \sqrt{\frac{2H^2\ln(4k^2SA/\delta)}{c\eta_k n_k(H-1)(k-1)}}\right) C\eta_k n_k (H-1)(k-1)$.

We denote event $\mathcal{E}_1$ as the events 1, 3, 4 hold simultaneously. We denote event $\mathcal{E}_2$ as the events 2, 3, 4 hold simultaneously.

Lemma E.7. *When for DP case, $\mathcal{E}_1$ holds (similar for EqOpt case, $\mathcal{E}_2$ holds), and when $\eta_k \leq \tilde{C}$, we have*

Proof. Recall

$$\mathcal{R}^*_h(\boldsymbol{p}^*, \boldsymbol{\pi}) = \sum_{\vartheta\in\{\alpha,\beta\}} p_{\vartheta} \cdot \mathbb{E}^{\pi^{*\vartheta},p^{*\vartheta}}[r^{*\vartheta}(s^{\vartheta}_h, a^{\vartheta}_h)].$$

Our aim is to bound

$$\mathcal{R}_{reg}(k) = \frac{1}{H}\sum_{h=1}^{H}\left(\mathcal{R}^*_h(\boldsymbol{p}^*,\boldsymbol{\pi}^*) - \mathcal{R}^*_h(\boldsymbol{p}^*,\boldsymbol{\pi}_k)\right) = \frac{1}{H}\sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right).$$

Let us first study $\sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right)$.

If $\{b^{\vartheta}_{k,h}\}_{k,h}$ are valid, by optimism, we have

$$\mathbb{E}\left[V^{\pi^{*\vartheta},p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right] \leq \mathbb{E}\left[V^{\pi^{*\vartheta},p^{\vartheta}_k}_{r^{\vartheta}_k}(s_1,1)\right].$$

As a result, we have

$$\sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right) \leq \sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{\vartheta}_k}_{r^{\vartheta}_k}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right).$$

Throughout the proof and proofs afterwards, let us take $\eta_k = k^{-1/3}$.

If $\eta_k \leq \tilde{C}$ (equivalently $k > (\tilde{C})^{-3}$ if we take $\eta_k = k^{-1/3}$), then by compatibility of $\{\hat{c}_{k,h}\}_{k,h}$ or $\{\hat{d}_{k,h}\}_{k,h}$, $(\pi^{*\alpha}, \pi^{*\beta})$ is a feasible solution to our algorithm, as a result

$$\begin{aligned}
\sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right) &\leq \sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{*\vartheta},p^{\vartheta}_k}_{r^{\vartheta}_k}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right) \\
&\leq \sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{\vartheta}_k}_{r^{\vartheta}_k}(s_1,1)\right] - \mathbb{E}\left[V^{\pi^{\vartheta}_k,p^{*\vartheta}}_{r^{*\vartheta}}(s_1,1)\right]\right) \\
&\leq \sum_{\vartheta} p^{\vartheta}\left[\mathcal{B}^{\pi^{\vartheta}_k,p^{\vartheta}_k}_{r^{\vartheta}_k}(s_{k,h},a_{k,h},h)\right].
\end{aligned}$$

Thus,

$$\begin{aligned}\mathcal{R}_{reg}(k) &= \frac{1}{H}\sum_{h=1}^{H}\left(\mathcal{R}_h^*(\boldsymbol{p}^*,\boldsymbol{\pi}^*) - \mathcal{R}_h^*(\boldsymbol{p}^*,\boldsymbol{\pi}_k)\right)\\ &= \frac{1}{H}\sum_{\vartheta} p^{\vartheta}\left(\mathbb{E}\left[V_{r^{*\vartheta}}^{\pi^{*\vartheta},p^{*\vartheta}}(s_1,1)\right] - \mathbb{E}\left[V_{r^{*\vartheta}}^{\pi_k^{\vartheta},p^{*\vartheta}}(s_1,1)\right]\right)\\ &\le \frac{1}{H}\sum_{\vartheta} p^{\vartheta}\left[\mathcal{B}_{r_k^{\vartheta}}^{\pi_k^{\vartheta},p_k^{\vartheta}}(s_{k,h},a_{k,h},h)\right].\end{aligned}$$

By Lemma B.4 of Brantley et al. (2020), with probability $1-2\delta$ for any $\vartheta \in \{\alpha,\beta\}$,

$$\left|\mathcal{B}_{r_k^{\vartheta}}^{\pi_k^{\vartheta},p_k^{\vartheta}}(s_{k,h},a_{k,h},h)\right| \le 4H^2\sqrt{\frac{2S\ln(16S^2AH^2k^3/\delta)}{\min_{s,a}N_k(s,a)}} + \frac{1}{k}.$$

Then, when $k \ge \frac{8H^2\ln(4k^2SA/\delta)}{C\eta_k n_k(H-1)}$, that is $k \ge C' + \left(\frac{8H^2\ln(4SA/\delta)}{Cn_k(H-1)}\right)^3$ for some constant C', we have

$$N_k^{\vartheta}(s,a) \ge \frac{1}{2}C\eta_k n_k(H-1)(k-1).$$

Thus,

$$\left|\mathcal{B}_{r_k^{\vartheta}}^{\pi_k^{\vartheta},p_k^{\vartheta}}(s_{k,h},a_{k,h},h)\right| \le 4H^2\sqrt{\frac{4S\ln(16S^2AH^2k^3/\delta)}{Ck^{-1/3}n_k(H-1)(k-1)}} + \frac{1}{k}.$$

Thus, with probability $1-2\delta$

$$\mathcal{R}_{reg}(k) \le 4H\sqrt{\frac{4S\ln(16S^2AH^2k^3/\delta)}{Ck^{-1/3}n_k(H-1)(k-1)}} + \frac{1}{kH}$$

□

Theorem E.2 (Restatement of Theorem 4.1)**.** *For type $\in \{DP, EqOpt\}$, with probability at least $1-\delta$, there exists a threshold $T = \mathcal{O}\left(\left(\frac{H\ln(SA/\delta)}{n_k}\right)^3\right)$, such that for all $k \ge T$,*

$$\mathcal{R}_{\text{reg}}^{type}(k) = \mathcal{O}\left(Hk^{-\frac{1}{3}}\sqrt{HS\ln(S^2AH^2k^3/\delta)}\right).$$

Proof. Noticing $n_k \ge 1$ and the result follows immediately from Lemma E.7. □

E.5 Proof of Theorem 4.2

Recall for the fairness constraints, we consider violation for each type of constraint in episode k as the following:

$$\mathcal{C}_{reg}^{DP}(k) = \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^{\alpha},p^{*\alpha}}(a_{k,h}^{\alpha}=1) - \mathbb{P}^{\pi_k^{\beta},p^{*\beta}}(a_{k,h}^{\beta}=1)\right|.$$

and

$$\mathcal{C}_{reg}^{EqOpt}(k) = \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^{\alpha},p^{*\alpha}}(a_{k,h}^{\alpha}=1|y_{k,h}^{\alpha}=1) - \mathbb{P}^{\pi_k^{\beta},p^{*\beta}}(a_{k,h}^{\beta}=1|y_{k,h}^{\beta}=1)\right|.$$

Theorem E.3 (Restatement of Theorem 4.2)**.** *For type* $\in \{DP, EqOpt\}$*, with probability at least* $1-\delta$*, there exists a threshold* $T = \mathcal{O}\left(\left(\frac{H\ln(SA/\delta)}{n_k}\right)^3\right)$*, such that for all* $k \geq T$*,*

$$\mathcal{C}_{\text{reg}}^{type}(k) \leq \mathcal{O}\left(k^{-\frac{1}{3}}\sqrt{SH\ln(S^2HAk^3/\delta)}\right).$$

Proof. Let us first consider $\mathcal{C}_{reg}^{DP}(k)$. Notice

$$\begin{aligned}
\mathcal{C}_{reg}^{DP}(k) &= \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1)\right| \\
&\leq \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\beta, p_k^\beta}(a_{k,h}^\beta = 1)\right| + \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1)\right| \\
&+ \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1) - \mathbb{P}^{\pi_k^\beta, p_k^\beta}(a_{k,h}^\beta = 1)\right| \\
&\leq \sum_h \frac{\hat{c}_{k,h}}{H} + \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\alpha, p_k^\alpha}(a_{k,h}^\alpha = 1)\right| \\
&+ \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1) - \mathbb{P}^{\pi_k^\beta, p_k^\beta}(a_{k,h}^\beta = 1)\right|.
\end{aligned}$$

Notice that switching $\pi^{*\vartheta}$ to π_k^ϑ doesn't change our argument in Lemma 4.2, thus with probability at least $1-\delta$, for ϵ_k (we take ϵ_k to be consistent to $\hat{c}_{k,h}$) for both $\vartheta = \{\alpha, \beta\}$

$$\left|\mathbb{P}^{\pi_k^\vartheta, p^{*\vartheta}}(a_{k,h}^\vartheta = 1) - \mathbb{P}^{\pi_k^\vartheta, p_k^\vartheta}(a_{k,h}^\vartheta = 1)\right| \leq H\sqrt{\frac{2S\ln(16SAk^2/(\epsilon_k\delta))}{\min_{s,a} N_k^\vartheta(s,a)}} + 2\epsilon_k HS.$$

Thus, with probability at least $1-2\delta$,

$$\mathcal{C}_{reg}^{DP}(k) \leq \sum_\vartheta 2H\sqrt{\frac{2S\ln(16SAk^2/(\epsilon_k\delta))}{\min_{s,a} N_k^\vartheta(s,a)}} + 4\epsilon_k HS.$$

By taking $\epsilon_k = 1/(kHS)$,

$$\mathcal{C}_{reg}^{DP}(k) \leq \sum_\vartheta 2H\sqrt{\frac{2S\ln(16S^2Ak^3H/(\delta))}{\min_{s,a} N_k^\vartheta(s,a)}} + \frac{4}{k}.$$

Now, let us consider

$$\mathcal{C}_{reg}^{EqOpt}(k) = \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha, p^{*\alpha}}(a_{k,h}^\alpha = 1 | y_{k,h}^\alpha = 1) - \mathbb{P}^{\pi_k^\beta, p^{*\beta}}(a_{k,h}^\beta = 1 | y_{k,h}^\beta = 1)\right|.$$

Similarly, by the triangle inequality,

$$\begin{aligned}\mathcal{C}_{reg}^{EqOpt}(k) &= \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha,p^{*\alpha}}(a_{k,h}^\alpha=1\big|y_{k,h}^\alpha=1)-\mathbb{P}^{\pi_k^\beta,p^{*\beta}}(a_{k,h}^\beta=1|y_{k,h}^\beta=1)\right|\\ &\le \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{P}^{\pi_k^\alpha,p_k^\alpha}(a_{k,h}^\alpha=1\big|y_{k,h}^\alpha=1)-\mathbb{P}^{\pi_k^\beta,p_k^\beta}(a_{k,h}^\beta=1|y_{k,h}^\beta=1)\right|\\ &+ \frac{1}{H}\sum_{h=1}^{H}\sum_{\vartheta}\left|\mathbb{P}^{\pi_k^\vartheta,p_k^\vartheta}(a_{k,h}^\vartheta=1\big|y_{k,h}^\vartheta=1)-\mathbb{P}^{\pi_k^\vartheta,p^{*\vartheta}}(a_{k,h}^\vartheta=1|y_{k,h}^\vartheta=1)\right|.\end{aligned}$$

Notice that switching $\pi^{*\vartheta}$ to π_k^ϑ doesn't change our argument in Lemma 4.3, thus with probability at least $1-2\delta$, for ϵ_k (we take ϵ_k to be consistent to $\hat{c}_{k,h}$),

$$\mathcal{C}_{reg}^{EqOpt}(k)\le\begin{cases}\sum_\vartheta \frac{6H\sqrt{\frac{2S\ln(32SAk^2/(\epsilon_k\delta))}{N_k^{\vartheta,min}}}+6\epsilon_k HS}{p_k^{\vartheta,min}\left(p_k^{\vartheta,min}-\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}}\right)}, & \text{if } p_k^{\vartheta,min}>\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k^{\vartheta,min}}};\\ 1, & \text{otherwise.}\end{cases}$$

Meanwhile, it also holds simultaneously that for all $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$|p_k(y=1|s,a)-p^*(y=1|s,a)|\le\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{N_k(s,a)}}.$$

Thus, if $4\sqrt{\frac{4\ln 2+2\ln(4SAk^2/\delta)}{\min_{s,a}N_k(s,a)}}<C$, we have

$$\mathcal{C}_{reg}^{EqOpt}(k)\le\sum_\vartheta\frac{6H\sqrt{\frac{2S\ln(32SAk^2/(\epsilon_k\delta))}{N_k^{\vartheta,min}}}+6\epsilon_k HS}{c^2/4}$$

Recall with probability at least $1-\delta$, for all $k\in[K]$, $(s,a)\in\mathcal{S}\times\mathcal{A}$, $\vartheta\in\{\alpha,\beta\}$

$$N_k^\vartheta(s,a)\ge\left(1-\sqrt{\frac{2H^2\ln(4k^2SA/\delta)}{C\eta_k n_k(H-1)(k-1)}}\right)C\eta_k n_k(H-1)(k-1).$$

Then, when $k\ge\frac{8H^2\ln(4k^2SA/\delta)}{C\eta_k n_k(H-1)}$,

$$N_k^\vartheta(s,a)\ge\frac{1}{2}C\eta_k n_k(H-1)(k-1).$$

Thus, we have the following properties:

- With probability $1-3\delta$, there exists a constant threshold C, for $k\ge C'+\left(\frac{8H^2\ln(4SA/\delta)}{Cn_k(H-1)}\right)^3$, we have

$$\min_{s,a}N_k^\vartheta(s,a)\ge\frac{1}{2}Cn_k(H-1)(k-1)k^{-1/3}.$$

  As a result,

$$\mathcal{C}_{reg}^{DP}(k)\le\sum_\vartheta 2H\sqrt{\frac{4S\ln(16S^2HAk^3/\delta)}{Cn_k(H-1)(k-1)k^{-1/3}}}+\frac{4}{k}.$$

- With probability $1-3\delta$, there exists a constant threshold C', for

$$k \geq \max\left\{C' + \left(\frac{8H^2\ln(4SA/\delta)}{Cn_k(H-1)}\right)^3, \left(\frac{32\ln 2 + 16\ln(4SA/\delta)}{C^3(H-1)n_k}\right)^3\right\},$$

$$\mathcal{C}_{reg}^{EqOpt}(k) \leq \sum_{\vartheta} \frac{24H\sqrt{\frac{4S\ln(32S^2HAk^3/\delta)}{Cn_k(H-1)(k-1)k^{-1/3}}} + 24/k}{C^2}.$$

Noticing $n_k \geq 1$, we obtain the final result. □

F FURTHER DETAILS ABOUT EXPERIMENTS

F.1 Additional Optimizations

For the formulation of our algorithm via occupancy measure, please refer to Appendix A. Here, we describe additional formulations for the surrogate optimization.

F.1.1 Demographic parity penalized objective surrogate

For episode k, the optimization problem can be reformulated as:

$$\max_{\rho} \sum_{x,y,a,h,\vartheta} p_\vartheta \rho_k^\vartheta(x,y,a,h)\hat{r}_k^\vartheta(x,y,a) - \lambda \sum_{h,a}(\sum_{y,x}\rho_k^\alpha(x,y,a,h) - \sum_{y,x}\rho_k^\beta(x,y,a,h))^2$$

such that

$$\forall \vartheta, x, t \ \ \frac{\rho_k^\vartheta(x,y=1,a=1,h)}{\sum_a \rho_k^\vartheta(x,y=1,a,h)} = \frac{\rho_k^\vartheta(x,y=0,a=1,h)}{\sum_a \rho_k^\vartheta(x,y=0,a,h)}$$

$$\forall \vartheta, x', y', h \quad \sum_a \rho_k^\vartheta(x',y',a,h+1) = \sum_{x,y,a} \rho_k^\vartheta(x,y,a,h) p_{(k)}^\vartheta(x',y'|x,y,a),$$

$$\forall \vartheta, x, y, a, h \ \ 0 \leq \rho_k^\vartheta(x,y,a,h) \leq 1, \quad \sum_{x,y,a} \rho_k^\vartheta(x,y,a,h) = 1.$$

F.1.2 Equal opportunity penalized objective surrogate

We make use of change of variable techniques to convert the polynomial optimization problem to a quadratic optimization problems for computational purposes. For episode k, the optimization problem can be reformulated as:

$$\max_{\rho,u,v} \sum_{x,y,a,h,\vartheta} p_\vartheta \rho_k^\vartheta(x,y,a,h)\hat{r}_k^\vartheta(x,y,a) - \sum_h \lambda(u_h - v_h)^2$$

such that

$$\forall h \ \ v_h = \sum_x \rho_k^\alpha(x,y=1,a=1,h) \sum_{x,a} \rho_k^\beta(x,y=1,a,h)$$

$$\forall h \ \ u_h = \sum_x \rho_k^\beta(x,y=1,a=1,h) \sum_{x,a} \rho_k^\alpha(x,y=1,a,h)$$

$$\forall \vartheta, x, t \ \ \frac{\rho_k^\vartheta(x,y=1,a=1,h)}{\sum_a \rho_k^\vartheta(x,y=1,a,h)} = \frac{\rho_k^\vartheta(x,y=0,a=1,h)}{\sum_a \rho_k^\vartheta(x,y=0,a,h)}$$

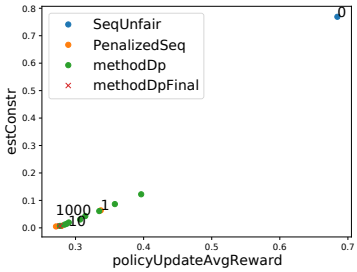
$$\forall \vartheta, x', y', h \quad \sum_a \rho_k^\vartheta(x',y',a,h+1) = \sum_{x,y,a} \rho_k^\vartheta(x,y,a,h) p_{(k)}^\vartheta(x',y'|x,y,a),$$

$$\forall \vartheta, x, y, a, h \ \ 0 \leq \rho_k^\vartheta(x,y,a,h) \leq 1, \quad \sum_{x,y,a} \rho_k^\vartheta(x,y,a,h) = 1.$$

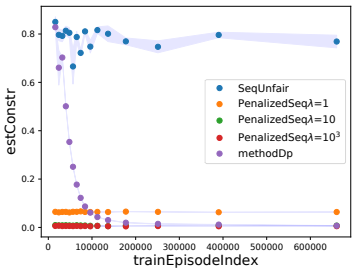
F.2 Additional Results for Synthetic Data

For the population dynamics, we model the initial qualification distribution as $\mathbb{P}^{\vartheta}(y_0^{\vartheta} = 1)$ and the initial feature distribution conditioned on qualification $\mathbb{P}^{\vartheta}(x_0 = j|y_0 = w)$. For a loan setting, we can interpret a higher feature value i as corresponding to a better credit score. Then, we generate the underlying group-independent and time invariant transition kernel $p^{*\vartheta}$ in the following way: we first set a distribution for $p^{*\vartheta}(y' = w'|y = w, a = v)$ for $w', w, v \in \{0, 1\}$. Then, we set $p^{*\vartheta}(x' = j'|x = j, y' = w', a = v)$ for $j', j \in \mathcal{X}$ and $w', v \in \{0, 1\}$. Thus we set $p^{*\vartheta}(x', y'|x, y, a)$ as $p^{*\vartheta}(y'|y, a)\ p^{*\vartheta}(x'|x, y', a)$

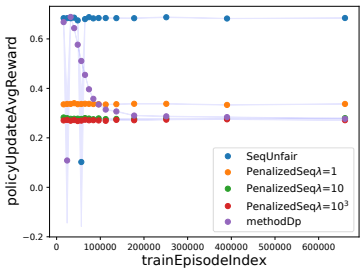
Figure 2a shows the Pareto frontier in terms of episodic total return and episodic step-average fairness violation for demographic parity, and Figure 2d shows the counterpart for equal opportunity. Figure 2b and 2c demonstrate the training dynamics of different algorithms for demographic parity, and Figures 2e and 2f demonstrate the counterpart for equal opportunity. Our proposed method converges to a stable level in terms of fairness violation over the training episodes. In addition, from the confidence intervals, we see that our algorithm has a much narrower confidence band than the baseline.



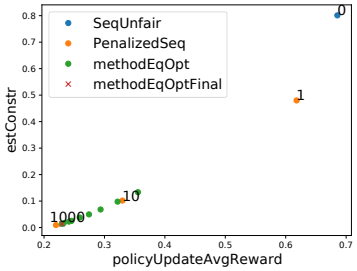
(a) DP Synthetic Pareto



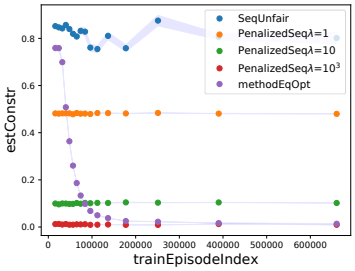
(b) DP Synthetic estConstr



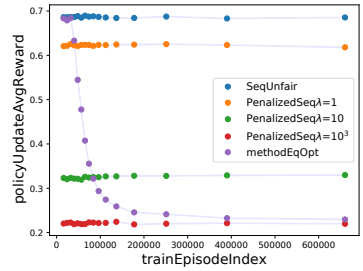
(c) DP Synthetic return



(d) EqOpt Synthetic Pareto



(e) EqOpt Synthetic estConstr



(f) EqOpt Synthetic return

Figure 2: Synthetic data result (methodDp and methodEqOpt are our proposed methods). 2a and 2d give the Pareto frontier, where the x axis is the average episodic return and the y axis is the constraint violation level. The cross marker gives the performance in the final episode and the text annotations give the penalty parameters. 2b and 2c give the constraint violation level during training. 2e and 2f give the average episodic return during training.

F.3 Detailed choice for parameters for experiments

We discuss the choice of parameters for our data generating processes.

F.3.1 Synthetic Data

We set $\mathbb{P}(x_0 = i|y_0 = 1) = 0.2$ and $\mathbb{P}(x_0 = i|y_0 = 0) = 0.2$ for initial probability conditioned on the qualification status of the individual.

We set $p^{*\vartheta}(y' = 1|y = 1, a = 1) = 0.6$, $p^{*\vartheta}(y' = 1|y = 1, a = 0) = 0.4$, $p^{*\vartheta}(y' = 1|y = 0, a = 1) = 0.6$ and $p^{*\vartheta}(y' = 1|y = 0, a = 0) = 0.4$.

We define $d_{\vartheta,j',j,w'} := p^{*\vartheta}(x' = j'|x = j, y' = w'), \forall \vartheta, j', j, w'$. We further define a vector: $D_{\vartheta,j,w'} = [d_{\vartheta,j',j,w'}]_{j'}$. And we set $D_{\vartheta,0,w'} = [0.3, 0.25, 0.2, 0.15, 0.1]$, $D_{\vartheta,1,w'} = [0.22, 0.26, 0.22, 0.17, 0.13]$, $D_{\vartheta,2,w'} = [0.17, 0.21, 0.24, 0.21, 0.17]$, $D_{\vartheta,3,w'} = [0.13, 0.17, 0.22, 0.26, 0.22]$ and $D_{\vartheta,4,w'} = [0.1, 0.15, 0.2, 0.25, 0.3], \forall \vartheta, w'$.

Here, we use an asymmetric $p^{*\vartheta}$ value, such that we have higher probability to obtain sampled individual who is qualified for the next time step if we give positive decision at this step. This reflects the fact that the sampled individual will be motivated by a positive decision from the decision maker and demotivated by a negative decision.

F.3.2 Semi-Realistic Data

Our experiments in the main context of FICO data make use of semi-realistic data, in that we need to define the population dynamics. For the full data generating process, we define $g_{\vartheta,j',j,1,1} := p^{*\vartheta}(x' = j'|x = j, y = 1, a = 1)$, $g_{\vartheta,j',j,1,0} := p^{*\vartheta}(x' = j'|x = j, y = 1, a = 0)$, $g_{\vartheta,j',j,0,1} := p^{*\vartheta}(x' = j'|x = j, y = 0, a = 1)$ and $g_{\vartheta,j',j,0,0} := p^{*\vartheta}(x' = j'|x = j, y = 0, a = 0)$. We further define a vector: $G_{\vartheta,j,w,v} = [g_{\vartheta,j',j,w,v}]_{j'}$, and we set, $G_{\vartheta,0,1,1} = [0.3, 0.25, 0.2, 0.15, 0.1]$, $G_{\vartheta,1,1,1} = [0.18, 0.27, 0.23, 0.18, 0.14]$, $G_{\vartheta,2,1,1} = [0.14, 0.18, 0.27, 0.23, 0.18]$, $G_{\vartheta,3,1,1} = [0.1, 0.15, 0.2, 0.3, 0.25]$, $G_{\vartheta,4,1,1} = [0.06, 0.13, 0.19, 0.24, 0.38]$, $G_{\vartheta,0,1,0} = [0.38, 0.24, 0.19, 0.13, 0.06]$, $G_{\vartheta,1,1,0} = [0.25, 0.3, 0.2, 0.15, 0.1]$, $G_{\vartheta,2,1,0} = [0.18, 0.23, 0.27, 0.18, 0.14]$, $G_{\vartheta,3,1,0} = [0.14, 0.18, 0.23, 0.27, 0.18]$, $G_{\vartheta,4,1,0} = [0.1, 0.15, 0.2, 0.25, 0.3]$, $G_{\vartheta,0,0,1} = [0.3, 0.25, 0.2, 0.15, 0.1]$, $G_{\vartheta,1,0,1} = [0.18, 0.27, 0.23, 0.18, 0.14]$, $G_{\vartheta,2,0,1} = [0.14, 0.18, 0.27, 0.23, 0.18]$, $G_{\vartheta,3,0,1} = [0.1, 0.15, 0.2, 0.3, 0.25]$, $G_{\vartheta,4,0,1} = [0.1, 0.15, 0.2, 0.25, 0.3]$, $G_{\vartheta,0,0,0} = [0.38, 0.24, 0.19, 0.13, 0.06]$, $G_{\vartheta,1,0,0} = [0.25, 0.3, 0.2, 0.15, 0.1]$, $G_{\vartheta,2,0,0} = [0.18, 0.23, 0.27, 0.18, 0.14]$, $G_{\vartheta,3,0,0} = [0.14, 0.18, 0.23, 0.27, 0.18]$, $G_{\vartheta,4,0,0} = [0.1, 0.15, 0.2, 0.25, 0.3]$.

Similar to the synthetic data generating process, $g_{\vartheta,j',j,w,v}$ are set such that x has a higher probability to transition to a higher value of x' for the next step when we make a positive decision at the current step, and a lower probability to transition to a lower value of x' for the next step when we make a negative decision at the current step.

For example, when $x = 2$, we have higher probability that x will transition to $x' = 3$ than $x' = 1$ given $a = 1$.