LONG-TERM FAIRNESS WITH UNKNOWN DYNAMICS

Tongxin Yin*
Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI 48109
tyin@umich.edu

Reilly Raab*
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
reilly@ucsc.edu

Mingyan Liu
Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI 48109
mingyan@umich.edu

Yang Liu
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yangliu@ucsc.edu

ABSTRACT

As populations adapt to algorithmic prediction, machine learning can myopically reinforce social inequalities or dynamically seek equitable outcomes. In this paper, we formalize prediction subject to long-term fairness as a constrained online reinforcement learning problem. This formulation can accommodate dynamical control objectives, such as inducing equitable population adaptations, that cannot be expressed by static formulations of fairness. By adapting recent work in online learning, we provide the first algorithm that guarantees simultaneous, probabilistic bounds on cumulative loss and cumulative violations of fairness (defined as statistical regularities between demographic groups) in this setting. We compare this algorithm to an off-the-shelf, deep reinforcement learning algorithm that lacks such safety guarantees, and to a repeatedly retrained, myopic classifier, as a baseline. We demonstrate that a reinforcement learning framework for long-term fairness allows algorithms to adapt to unknown dynamics and sacrifice short-term profit or fairness to drive a classifier-population system towards more desirable equilibria. Our experiments model human populations according to evolutionary game theory, using real-world data to set an initial state.

1 INTRODUCTION

As machine learning (ML) algorithms are deployed for tasks with real-world social consequences (e.g., school admissions, loan approval, medical interventions, etc.), the possibility exists for runaway social inequalities (Crawford & Calo, 2016; Chaney et al., 2018; Fuster et al., 2018; Ensign et al., 2018). While "fairness" has become a salient ethical concern in contemporary research, the closed-loop dynamics of real-world systems comprising ML policies and populations that mutually adapt to each other (Fig. 1 in the supplementary material) remain poorly understood.

Our primary contribution is to consider the problem of *long-term fairness*, or algorithmic fairness in the context of a dynamically responsive population, as a reinforcement learning (RL) problem subject to constraint. The central learning task is to develop a policy that minimizes cumulative loss (e.g., financial risk, negative educational outcomes, misdiagnoses, etc.) incurred by an ML agent interacting with a human population up to a finite time horizon, subject to constraints on cumulative "violations of fairness", which we refer to in a single time step as *disparity* and cumulatively as *distortion*.

Our central hypothesis is that an RL formulation of long-term fairness can allow an agent to learn to **sacrifice short-term utility in order to drive the system towards more desirable equilibria**. The core practical difficulties posed by our general problem formulation, however, are the potentially unknown dynamics of the system under control, which must be determined by the RL agent *online*

*These authors contributed equally to this work.

(i.e., during actual deployment), and the general non-convexity of the losses or constraints considered. Additionally, we address continuous state and action spaces, in general, which preclude familiar methods with performance guarantees in discrete settings.

Our secondary contributions are 1) to show that long-term fairness can be solved within asymptotic, probabilistic bounds under certain dynamical assumptions and 2) to demonstrate that the problem of long-term fairness can also be addressed more flexibly . For theoretical guarantees, we develop `L-UCBFair`, an online RL method, and prove sublinear bounds on regret (suboptimality of cumulative loss) and distortion (suboptimality of cumulative disparity) with high probability (Section 3.1). To demonstrate practical solutions, we consider a time-dependent Lagrangian relaxation of the fairness constraint using well-known deep reinforcement learning method (viz., TD3), an approach we abbreviate as `R-TD3`. We compare `L-UCBFair` and `R-TD3` to a baseline, myopic policy in interaction with simulated populations initialized with synthetic or real-world data and updated according to evolutionary game theory (Appendix A).

Throughout, we consider fairness in terms of statistical regularities across (ideally) socioculturally meaningful *groups*. Acknowledging that internal conflict exists between different statistical measures of fairness, we show that an RL approach to long-term fairness can mitigate trade-offs between fairness defined on the statistics of immediate policy decision *outcomes* (Chen et al., 2022), (e.g., acceptance rate disparities (Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015)) and underlying distributional parameters (e.g., qualification rate (Raab & Liu, 2021; Zhang et al., 2020)).

1.1 Related Work

Our effort to formalize long-term fairness as a reinforcement learning problem bridges recent work on "fairness in machine learning", which has developed in response to the proliferation of data-driven methods in society, and "safe reinforcement learning", which seeks theoretical safety guarantees in the control of dynamical systems.

Dynamics of Fairness in Machine Learning We distinguish long-term fairness from the dynamics of fair allocation problems (Joseph et al., 2016; Jabbari et al., 2017; Tang et al., 2021; Liu et al., 2017) and emphasize side-effects of algorithmic decisions affecting future decision problems. By formalizing long-term fairness in terms of cumulative losses and disparities, we iterate on a developing research trend that accounts for the dynamical response of a human population to deployed algorithmic prediction: both as a singular reaction (Liu et al., 2018; Hu et al., 2019; Perdomo et al., 2020) or as a sequence of mutual updates to the population and the algorithm (Coate & Loury, 1993; D'Amour et al., 2020; Zhang et al., 2020; Heidari et al., 2019; Wen et al., 2019; Liu et al., 2020; Hu & Chen, 2018; Mouzannar et al., 2019; Williams & Kolter, 2019; Raab & Liu, 2021). In particular, Perdomo et al. (2020) introduces the concept of "performative prediction", analyzing the fixed points of interactions between a population and an algorithmic classifier, but with state treated as a pure function of a classifier's actions. For more realistic dynamics, Mouzannar et al. (2019) and Raab & Liu (2021) model updates to qualification rates that depend on both previous state and the classifier's actions, but only treat myopic classifiers that optimize immediate utility (subject to fairness constraints) rather than learning to anticipate dynamical population responses.

Safe Reinforcement Learning `L-UCBFair` furthers recent efforts in safe RL. While "model-based" approaches, in which the algorithm learns an explicit dynamical model of the environment, constitute one thread of prior work (Efroni et al., 2020; Singh et al., 2020; Brantley et al., 2020; Zheng & Ratliff, 2020; Kalagarla et al., 2021; Liu et al., 2021; Ding et al., 2021), such algorithms are typified by significant time and space complexity. Among "model-free" algorithms, the unknown dynamics of our setting preclude the use of a simulator that can generate arbitrary state-action pairs Xu et al. (2021); Ding et al. (2020); Bai et al. (2022). While Wei et al. (2022) introduce a model-free and simulator-free algorithm, the tabular setting considered is only applicable to discrete state and action spaces. To tackle continuous state space, Ding et al. (2021); Ghosh et al. (2022) consider linear dynamics: Ding et al. (2021) develop a primal-dual algorithm with safe exploration, and Ghosh et al. (2022) use a softmax policy design. Both algorithms are based on the work of Jin et al. (2020), which proposed a least squares value iteration method, using an Upper Confidence Bound (UCB) (Auer et al., 2002) to estimate a state-action "Q" function. To our knowledge, `L-UCBFair` is the first model-free, simulator-free RL algorithm that provides theoretical safety guarantees for both discrete

and **continuous** state and action spaces. Moreover, `L-UCBFair` achieves bounds on regret and distortion as tight as any algorithm thus far with discrete action space (Ghosh et al., 2022).

2 Problem Formulation

Consider a binary classification task as starting point for our formulation, though the formal problem we propose is more widely applicable. To this initial task, we introduce "fairness" constraints, then population dynamics, and then cumulative loss and "disparity", before formalizing the problem of optimizing cumulative loss subject to constraints on cumulative disparity.

We introduce the following notation: a random individual, sampled i.i.d. from a population, has features $X \in \mathbf{R}^d$, a label $Y \in \{-1, 1\}$, and a demographic group $G \in \mathcal{G}$ (where $\mathcal{G} = [n]$ for $n \geq 2$). Denote the joint distribution of these variables in the population as $s \coloneqq \Pr(X, Y, G)$. The task is to predict Y (as $\hat{Y}$) from X and G. Specifically, the task is to choose a classifier a, such that $\hat{Y} \sim a(X, G)$, that minimizes some bounded loss $\mathscr{L} \in [0, 1]$ over s. This **basic classification task** is $\min_a \mathscr{L}(s, a)$. In general, we allow arbitrary, (unit-interval) bounded loss functions $\mathscr{L}$, though, typically, $\mathscr{L}$ corresponds to the expectation value of a loss function L defined for individuals drawn from s, such as such as zero-one-loss: $\mathscr{L}(s, a) \overset{e.g.}{=} \mathrm{E}_{\substack{X,Y,G\sim s \\ \hat{Y}\sim a(X,G)}} \left[L(Y, \hat{Y})\right]$.

The standard **"fair" classification task** (*without* a dynamically responsive population) is to constrain classifier a such that the *disparity* $\mathscr{D} \in [0, 1]$ induced on distribution s by a is bounded by some value $c \in [0, 1]$. That is, $\min_a \mathscr{L}(s, a) \mathrm{subject to} \mathscr{D}(s, a) \leq c$. A standard example of disparity is the expected divergence of group acceptance rates β, which is consistent with enforcing "demographic parity" Dwork et al. (2012). For example, when $\mathcal{G} = \{g_1, g_2\}$,

$$\mathscr{D}(s, a) \overset{e.g.}{=} \left|\beta_{s,a}(g_1) - \beta_{s,a}(g_2)\right|^2, \quad \text{where } \beta_{s,a}(g) \coloneqq \Pr_{\substack{X,Y,G\sim s \\ \hat{Y}\sim a(X,G)}} (\hat{Y}{=}1 \mid G{=}g).$$

We also consider measures of fairness based on inherent population statistics (e.g., parity of group qualification rates $\Pr(Y{=}1 \mid G{=}g)$), which must be driven dynamically Raab & Liu (2021); Zhang et al. (2020). Such notions of disparity are well-suited to an RL formulation of long-term fairness.

State, Action, and Policy For iterated classification tasks, we identify the distribution $s \in \mathcal{S}$ of individuals in the population as a *state* and the classifier $a \in \mathcal{A}$ as an *action*. While state space $\mathcal{S}$ may encompass arbitrary distributions, we assume that action space $\mathcal{A}$ admits a Euclidean metric, under which it is closed (i.e., $\mathcal{A}$ is isomorphic to $[0, 1]^m, m \in \mathbf{Z}_{>0}$). At a given time τ, a_τ is sampled stochastically according to the current *policy* π_τ: $a_\tau \sim \pi_\tau(s_\tau)$. We assume s_τ is fully observable at time τ. In practice, s_τ must be approximated from finitely many empirical samples, though this caveat introduces well-understood errors that vanish in the limit of infinitely many samples.

Dynamics In contrast to a "one-shot" fair classification task, we assume that a population may react to classification, inducing the distribution s to change. Importantly, such "distribution shift" is a well-known, real-world phenomenon that can increase realized loss and disparity when deployed classification policies are fixed Chen et al. (2022). For classification policies that free to change in response to a mutating distribution s, subsequent classification tasks depend on the (stochastic) predictions made in previous tasks. In our formulation, we assume the existence of dynamical kernel $\mathbf{P}$ that maps a state s and action a at time τ to a *distribution over* possible states at time $\tau + 1$, $s_{\tau+1} \sim \mathbf{P}(s_\tau, a_\tau)$. We stipulate that $\mathbf{P}$ may be initially unknown, but it does not explicitly depend on time and may be reasonably approximated "online". While real-world dynamics may depend on information other than the current distribution $\Pr(X, Y, G)$ (e.g., exogenous parameters, history, or additional variables of state), we identify s with the current distribution for simplicity. Fig. 1 provides a conceptual, graphical depiction of a population's response to deployed algorithmic policy, effecting a transition of state s.

Reward and Utility, Value and Quality Functions Because standard RL literature motivates *maximizing reward* rather than *minimizing loss*, let us define the instantaneous reward $r \in [0, 1]$ and a separate, instantaneous "utility" $g \in [0, 1]$ for an RL agent as $r(s_\tau, a_\tau) \coloneqq 1 - \mathscr{L}(s_\tau, a_\tau),\ g(s_\tau, a_\tau) \coloneqq 1 - \mathscr{D}(s_\tau, a_\tau)$, where r and g do not explicitly depend on time τ. Learnable dynamics inspire us to optimize anticipated *cumulative* reward, given constraints on anticipated *cumulative* utility. Let j represent either reward r or utility g. We use the letter V (for "value") to denote the future expected

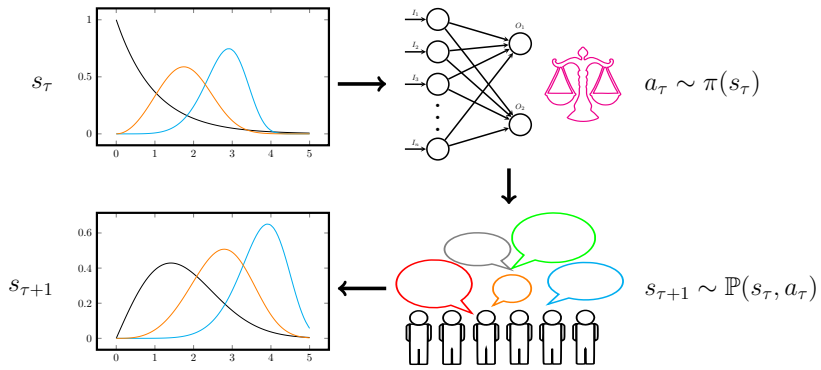


Figure 1: The interaction of an algorithmic classifier and a reactive population. Given state s_τ, the classifier uses policy π to select action a_τ. The population, in state s_τ, reacts to a_τ , transitioning state to $s_{\tau+1}$, then the process repeats.

accumulation of j over steps $[h, ..., H]$ (without time-discounting) starting from state s, using policy π. Likewise, we denote the "quality" of an action a in state s with the letter Q. For $j \in \{r, g\}$, $V^\pi_{j,h}(s) \coloneqq \mathrm{E}\left[\sum_{\tau=h}^{H} j(s_\tau, a_\tau) | s_h = s\right]$, $Q^\pi_{j,h}(s,a) \coloneqq \mathrm{E}\left[\sum_{\tau=h}^{H} j(s_\tau, a_\tau)) \mid s_h = s, a_h = a\right]$. By the boundedness of $r, g \in [0, 1]$, V and Q belong to the interval $[0, H - h + 1]$.

The central problem explored in this paper is

$$\max_{\pi} \quad V^\pi_{r,1}(s) \quad \text{subject to} \quad V^\pi_{g,1}(s) \geq \tilde{c} \tag{1}$$

We emphasize that this construction of long-term fairness considers a finite time horizon of H steps and denote the optimal value of π as $\pi^\star$.

The Online Setting In the online setting, learning dynamics is only possible through actual deployments of policy. As it is not possible to unconditionally guarantee constraint satisfaction in Eq. (1) over a finite number of episodes, we instead measure two types of *regret*: one that measures the suboptimality of a policy with respect to cumulative incurred loss, which we will continue to call "regret", and one that measures the suboptimality of a policy with respect to cumulative induced disparity, which we will call "distortion". Note that we define regret and distortion in Eq. (2) by marginalizing over the stochasticity of state transitions and the sampling of actions:

$$\mathrm{Regret}(\pi, s_1) \coloneqq V^{\pi^*}_{r,1}(s_1) - V^\pi_{r,1}(s_1), \ \mathrm{Distortion}(\pi, s_1) \coloneqq \max\left[0, \tilde{c} - V^\pi_{g,1}(s_1)\right] \tag{2}$$

3 Algorithms and Analysis

It is possible to provide guarantees for long-term fairness in the online setting: We develop `L-UCBFair`, the first model-free algorithm to provide such guarantees with continuous state and action spaces, and prove probabilistic, sublinear bounds for regret and distortion under appropriate assumptions and parameters (Appendix C, Appendix D.1).

3.1 L-UCBFair

Episodic MDP `L-UCBFair` inherits from a family of algorithms that treat an episodic Markov decision process (MDP) (Jin et al., 2020). We first map the long-term fairness problem to MDP$(\mathcal{S}, \mathcal{A}, H, \mathbf{P}, \mathscr{L}, \mathscr{D})$. The algorithm runs for K *episodes*, each consisting of H time steps. At the beginning of each episode, which we index with k, the agent commits to a sequence of policies $\pi^k = (\pi^k_1, \pi^k_2, ..., \pi^k_H)$ for the next H steps. At each step h within an episode, an action $a^k_h \in \mathcal{A}$ is sampled according to policy π^k_h, then the state s^k_{h+1} is sampled according to the transition kernel $\mathbf{P}(s^k_h, a^k_h)$. s^k_1 is sampled arbitrarily with each episode.

Algorithm 1 L-UCBFair

Input: A set of points $\{I_0, I_1, \cdots, I_M\}$ satisty Definition 3.1. $\epsilon_I = \frac{1}{2\rho(1+\chi)KH}$.
ν_1=0. $w_{r,h}$=$w_{g,h}$=0. α=$\frac{\log(M)K}{2(1+\chi+H)}$. η=$\chi/\sqrt{KH^2}$. β=$C_1 dH\sqrt{\log(4\log MdT/p)}$, $\varsigma = 1$.
for episode $k = 1, 2, ..., K$ **do**
  Receive the initial state s_1^k.
  for step $h = H, H-1, \cdots, 1$ **do**
    $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi\left(s_h^\tau, a_h^\tau\right) \phi\left(s_h^\tau, a_h^\tau\right)^T + \varsigma \mathbf{I}$
    for $j \in \{r, g\}$ **do**
      $w_{j,h}^k \leftarrow \left(\Lambda_h^k\right)^{-1} \left[\sum_{\tau=1}^{k-1} \phi\left(s_h^\tau, a_h^\tau\right)\left(j\left(s_h^\tau, a_h^\tau\right) + V_{j,h+1}^k\left(s_{h+1}^\tau\right)\right)\right]$
    end for
    for iteration $i = 1, \cdots, M$ and index $j \in \{r, g\}$ **do**
      $\xi_{i,j} \leftarrow \left(\phi(\cdot, I_i)^T \left(\Lambda_h^k\right)^{-1} \phi(\cdot, I_i)\right)^{1/2}$, $\quad Q_{j,h}^k(\cdot, I_i) \leftarrow \min\left[\left\langle w_{j,h}^k, \phi(\cdot, I_i)\right\rangle + \beta\xi_{i,j}, H\right]$
    end for
    $\mathrm{SM}_{h,k}(I_i \mid \cdot) = \frac{\exp\left(\alpha\left(Q_{r,h}^k(\cdot, I_i) + \nu_k Q_{g,h}^k(\cdot, I_i)\right)\right)}{\sum_j \exp\left(\alpha\left(Q_{r,h}^k(\cdot, I_j) + \nu_k Q_{a,h}^k(\cdot, I_j)\right)\right)}$
    $\pi_h^k(a \mid \cdot) \leftarrow \frac{1}{\int_{b \in \mathcal{I}(a)} db} \mathrm{SM}_{h,k}(I(a) \mid \cdot)$
    $V_{r,h}^k(\cdot) \leftarrow \int_{a \in \mathcal{A}} \pi_h^k(a \mid \cdot) Q_{r,h}^k(\cdot, a) da, \quad V_{g,h}^k(\cdot) \leftarrow \int_{a \in \mathcal{A}} \pi_h^k(a \mid \cdot) Q_{g,h}^k(\cdot, a) da$
  end for
  for step $h = 1, \cdots, H$ **do**
    Compute $Q_{r,h}^k\left(s_h^k, I_i\right), Q_{g,h}^k\left(s_h^k, I_i\right), \pi\left(I_i \mid s_h^k\right)$.
    Take action $a_h^k \sim \pi_h^k\left(\cdot \mid s_h^k\right)$ and observe s_{h+1}^k.
  end for
  $\nu_{k+1} = \max\left\{\min\left\{\nu_k + \eta\left(\tilde{c} - V_{g,1}^k\left(s_1\right)\right), \mathscr{V}\right\}, 0\right\}$
end for

Episodic Regret and The Lagrangian Because L-UCBFair predetermines its policy for an entire episode, we amend our definition of regret and distortion over HK time steps as a sum over K episodes of length H. $\mathrm{Regret}(K) = \sum_{k=1}^K \left(V_{r,1}^{\pi^*}\left(s_1^k\right) - V_{r,1}^{\pi^k}\left(s_1^k\right)\right)$, $\mathrm{Distortion}(K) = \max\left[0, \sum_{k=1}^K \left(\tilde{c} - V_{g,1}^{\pi^k}\left(s_1^k\right)\right)\right]$. For the Lagrangian $\mathcal{L}(\pi, \nu) := V_{r,1}^\pi(s) + \nu\left(V_{g,1}^\pi(s) - \tilde{c}\right)$ associated with Eq. (1), with dual variable $\nu \geq 0$, L-UCBFair approximately solves the primal problem $\max_\pi \min_\nu \mathcal{L}(\pi, \nu)$, which is non-trivial, since the objective function is seldom concave in practical parameterizations of π. Let ν^* to denote the optimal value of ν.

3.1.1 Explicit Construction

L-UCBFair, or "LSVI-UCB for Fairness " (Algorithm 1) is based on an optimistic modification of least-squares value iteration (LSVI), where optimism is realized by an upper-confidence bound (UCB), as in LSVI-UCB (Jin et al., 2020). For each H-step episode k, L-UCBFair maintains estimates for Q_r^k, Q_g^k and a time-indexed policy π^k. L-UCBFair updates Q_r^k, Q_g^k, interacts with the environment, and updates the dual variable ν_k (constant in k).

LSVI-UCB (Jin et al., 2020) The estimation of Q is challenging, as it is impossible to iterate over all s, a pairs when $\mathcal{S}$ and $\mathcal{A}$ are continuous and $\mathbf{P}$ is unknown. LSVI parameterizes $Q_h^\star(s, a)$ by the linear form $\mathrm{w}_h^\top \phi(s, a)$, as used in Jin et al. (2020), and updates

$$\mathbf{w}_h \leftarrow \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\tau=1}^{k-1} \left[r_h\left(s_h^\tau, a_h^\tau\right) + \max_{a \in \mathcal{A}} Q_{h+1}\left(s_{h+1}^\tau, a\right) - \mathbf{w}^\top \phi\left(s_h^\tau, a_h^\tau\right)\right]^2 + \varsigma \|\mathbf{w}\|^2.$$

In addition, a "bonus term" $\beta\left(\phi^\top \Lambda_h^{-1} \phi\right)^{1/2}$ is added to the estimate of Q to encourage exploration.

Adaptive Search Policy Unlike Ding et al. (2021) and Ghosh et al. (2022), we assume a continuous action space $\mathcal{A}$, which renders the independent computation of Q_r^k, Q_g^k for each action impossible. To handle this issue, we propose an adaptive search policy, sampling from finitely many Voronoi

regions of action space with a softmax scheme, then sampling an action uniformly at random from the selected partition.

Definition 3.1. Given a set of distinct actions $I = \{I_0, \cdots, I_M\} \subset \mathcal{A}$, where $\mathcal{A}$ is a closed set in Euclidean space, define $\mathcal{I}_i = \{a \colon \|a - I_i\|_2 \leq \|a - I_j\|_2, \forall i < j\}$ as the subset of actions closer to I_i than to I_j, i.e., the Voronoi region corresponding to locus I_i, with tie-breaking imposed by the order of indices i. Also define the locus function $I(a) = \min_i \operatorname{argmin}_{I_i} \|a - I_i\|_2$.

Assumption 3.2. There exists $\rho > 0$, such that $\|\phi(s, a) - \phi(s, a')\|_2 \leq \rho \|a - a'\|_2$.

Assumption 3.2 bounds the difference in the estimated quality of action-value pairs for nearby actions.

For `L-UCBFair`, the update method for the dual variable ν in the Lagrangian is also essential. Since $V^{\pi}_{r,1}(s)$ and $V^{\pi}_{g,1}(s)$ are unknown, we use $V^{k}_{r,1}(s)$ and $V^{k}_{g,1}(s)$ to estimate them. ν is iteratively updated by minimizing the Lagrangian with step-size η, and $\mathscr{V}$ is an upper bound for ν (Assumption C.2). A similar method is also used in Ding et al. (2021); Ghosh et al. (2022).

Theorem 3.3 (Boundedness)**.** *With probability $1 - p$, there exists a constant b such that `L-UCBFair` (Algorithm 1) achieves Regret$(K) = \tilde{\mathcal{O}}\left(H^2\sqrt{d^3K}\right)$, Distortion $(K) = \tilde{\mathcal{O}}\left(H^2\sqrt{d^3K}\right)$.*

Compared to the algorithms introduced by Ding et al. (2021); Ghosh et al. (2022), which work with discrete action space, `L-UCBFair` guarantees the same asymptotic bounds on regret and distortion (Appendix D.1).

3.2 `R-TD3`

Technical assumptions that support `L-UCBFair` (Appendix C) are often violated in practice. We therefore consider more flexible reinforcement learning methods on a (Lagrangian) relaxation of the long-term fairness problem, $\min_\pi \mathrm{E}_{a_\tau \sim \pi(s_\tau)} \left[\sum_{\tau=1}^{H} \left[\kappa_\tau \mathscr{L}(s_\tau, a_\tau) + \lambda_\tau \mathscr{D}(s_\tau, a_\tau)\right]\right]$, where $s_{\tau+1} \sim \mathbf{P}(s_\tau, a_\tau), \lambda_\tau = \tau/H$, and $\kappa_\tau = 1 - \lambda_\tau$. Specifically, we experiment with "Twin-Delayed Deep Deterministic Policy Gradient" (TD3) (Fujimoto et al., 2018) with the implementation and default parameters provided by the open-source package "Stable Baselines 3" (Raffin et al., 2021) Strictly applied, myopic fairness constraints can lead to undesirable dynamics and equilibria (Raab & Liu, 2021). Relaxing these constraints (hard $\rightarrow$ soft) for the near future while emphasizing them long-term, we demonstrate classifiers that learn to transition to more favorable equilibria.

3.3 Experiments

We conduct extensive experiments to compare the performance of `L-UCBFair` and `R-TD3` to a baseline agent in Appendix A and Appendix B. We demonstrate that desirable social outcomes that are in conflict with myopic optimization may be realized using a reinforcement learning formalism of long-term fairness. In addition, we demonstrate that definitions of fairness that may be mutually incompatible for an unchanging population — such as parity in qualification rates and acceptance rates across groups — can be reconciled in the long-term framing, where the dynamic response of a population provides additional freedom.

4 Conclusion

Our work frames long-term fairness as an online reinforcement learning problem. We have shown that this problem 1) admits solutions with theoretical guarantees and 2) can be relaxed to accommodate a wider class of recent advances in reinforcement learning. Our experiments demonstrate that tensions between different notions of fairness, such as acceptance rate and qualification rate parity across groups, can be resolved when a policy learns to sacrifice short-term utility or fairness to induce dynamics resulting in more favorable long-term equilibria. We hope our contributions spur interest in long-term mechanisms and incentive structures for machine learning to be a driver of positive social change.

REFERENCES

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3682–3689, 2022.

Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 224–232. ACM, 2018.

Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *arXiv preprint arXiv:2206.00129*, 2022.

Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pp. 1220–1240, 1993.

Kate Crawford and Ryan Calo. There is a blind spot in AI research. *Nature News*, 538(7625):311, 2016.

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*, pp. 525–534, 2020.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference of Fairness, Accountability, and Transparency*, 2018.

Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, and Emmanouil Zampetakis. Optimal approximation-smoothness tradeoffs for soft-max functions. *Advances in Neural Information Processing Systems*, 33:2651–2660, 2020.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? The effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets*, 2018.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *arXiv preprint arXiv:2206.11889*, 2022.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *the International Conference on Machine Learning (ICML)*, 2019.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268, 2019.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International conference on machine learning*, pp. 1617–1626. PMLR, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8030–8037, 2021.

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

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368. ACM, 2019.

Ling Pan, Qingpeng Cai, Qi Meng, Wei Chen, Longbo Huang, and Tie-Yan Liu. Reinforcement learning with dynamic boltzmann softmax updates. *arXiv preprint arXiv:1903.05926*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34:26053–26065, 2021.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Wei Tang, Chien-Ju Ho, and Yang Liu. Bandit learning with delayed impact of actions. *Advances in Neural Information Processing Systems*, 34:26804–26817, 2021.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3274–3307. PMLR, 2022.

Min Wen, Osbert Bastani, and Ufuk Topcu. Fairness with dynamics. *arXiv preprint arXiv:1901.08568*, 2019.

Joshua Williams and J Zico Kolter. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837*, 2019.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *arXiv preprint arXiv:2010.11300*, 2020.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pp. 620–629. PMLR, 2020.

A Simulated Environments

A.1 Greedy Baseline

In our experiments, we compare `L-UCBFair` and `R-TD3` to a "Greedy Baseline" agent as a proxy for a myopic status quo in which policy is repeatedly determined by optimizing for immediate utility, without regard for the population dynamics induced by algorithmic actions. Our chosen algorithm for the greedy baseline is simply gradient descent in f, defined as loss regularized by disparity, performed anew with each time step with fixed parameter λ.

$$f_\tau(\pi) = \mathop{\mathrm{E}}_{a_\tau \sim \pi}[(1-\lambda)\mathscr{L}(s_\tau, a_\tau) + \lambda \mathscr{D}(s_\tau, a_\tau)] \tag{3}$$

While such an algorithm does not guarantee constraint satisfaction, it is nonetheless "constraint aware" in precisely the same way as a firm that (probabilistically) incurs penalties for violating constraints.

A.2 Setting

We describe our experiments with the algorithms we have detailed for long-term fairness as an RL problem: We consider a series of binary ($Y \in \{-1, 1\}$ classification tasks on a population of two groups $\mathcal{G} = \{g_1, g_2\}$ modeled according to evolutionary game theory (using replicator dynamics). We consider two families of distributions of real-valued features for the population: One that is purely synthetic, for which $X \sim \mathcal{N}(Y, 1)$, independent of group G, and one that is based on a logistic regression to real-world data. Both families of distributions are parameterized by the joint distribution $\Pr(Y, G)$. RL agents are trained on episodes of length H initialized with randomly sampled states.

The following assumptions simplify our hypothesis space for classifiers in order to better handle continuous state space. These assumptions appeared in Raab & Liu (2021).

Assumption A.1 (Well-behaved feature)**.** For purely synthetic data, we require X to be a "well-behaved" real-valued feature or "score" within each group. That is,

$$\forall g, \quad \Pr(Y{=}1 \mid G{=}g, X{=}x) \text{ strictly increases in } x$$

As an intuitive example of Assumption A.1, if Y represents qualification for a fixed loan and X represents credit-score, we require higher credit scores to strongly imply higher likelihood that an individual is qualified for the loan.

Theorem A.2 (Threshold Bayes-optimality)**.** *For each group g, when Assumption A.1 is satisfied, the Bayes-optimal, deterministic binary classifier is a threshold policy*

$$\hat{Y} = 1 \text{ if } x \geq A_g \text{ and } -1 \text{ otherwise}$$

where A_g is the feature threshold for group g.

As a result of Theorem A.2, we consider our action space to be the space of group-specific thresholds, and denote an individual action as the vector $\mathbf{A} \coloneqq (A_1, A_2, ..., A_n)$.

A.3 Replicator Dynamics

Our use of replicator dynamics closely mirrors that of Raab & Liu (2021) as an "equitable" model of a population, in which individuals my be modeled identically, independently of group membership, yet persistent outcome disparities may nonetheless emerge from disparate initial conditions between groups. In particular, we parameterize the evolving distribution $\Pr(X, Y \mid G)$, assuming constant group sizes, in terms of "qualification rates" $q_g \coloneqq \Pr(Y{=}1 \mid G{=}g)$ and update these qualification rates according to the discrete-time replicator dynamics:

$$q_g[t+1] = q_g[t]\frac{W_1^g[t]}{\overline{W}^g[t]}; \quad \overline{W}^g[t] \coloneqq W_1 q_g + (1-q_g)W_{-1}$$

In this model, the *fitness* $W_y^g > 0$ of label $Y{=}y$ in group $G{=}g$ may be interpreted as the "average utility to the individual" in group g of possessing label y, and thus relative *replication rate* of label y in group g, as agents update their labels by mimicking the successful strategies of in-group peers.

Also following Raab & Liu (2021), we model W_y^g in terms of acceptance and rejection rates with a group-independent utility matrix U:

$$W_y^g = \sum_{\hat{y} \in \{-1,1\}} U_{y,\hat{y}} \Pr\Big(\hat{Y}{=}\hat{y} \mid Y{=}y, G{=}g\Big)$$

We choose the matrix U to eliminate dominant strategies (i.e., agents prefer one label over another, independent of classification), assert that agents always prefer acceptance over rejection, and to imply that the costs of qualification are greater than the costs of non-qualification among accepted individuals. While other parameterizations of U are valid, this choice of parameters guarantees internal equilibrium of the replicator dynamics for a Bayes-optimal classifier and "well-behaved" scalar-valued feature X, such that $\Pr(Y{=}1 \mid X{=}x)$ is monotonically increasing in x (Raab & Liu, 2021).

A.4 Data Synthesis and Processing

In addition to a synthetic distribution, for which we assume $X \sim \mathcal{N}(Y, 1)$, independent of G, for all time, we also consider real-world distributions in simulating and comparing algorithms for "long-term fairness". In both cases, as mentioned above, we wish to parameterize distributions in terms of qualification rates q_g. As we perform binary classification on discrete groups and scalar-valued features, in addition to parameterizing a distribution in terms of q_g, we desire a scalar-valued feature for each example, rather than the multi-dimensional features common to real-world data. Our solution to parameterize a distribution of groups and scalar features is to use an additional learning step for "preprocessing": Given a static dataset $\mathcal{D}$ from which (X', Y, G) is drawn i.i.d., (e.g., the "Adult Data Set" Dua & Graff (2017)), at each time-step, we train a stochastic binary classifier $\tilde{a}$, such that $\hat{Y}' \sim \tilde{a}(X', G)$ with a loss that re-weights examples by label value, in order to simulate the desired q_g: $\min_{\tilde{a}} \quad \mathrm{E}_{\tilde{a},\mathcal{D}}[w(X', Y, G)L(Y, \hat{Y}')]$, where $w(X', Y, G) = [(1-Y)/2 + Yq_g]/\,\mathrm{E}_{\mathcal{D}}[Y|G]$, L is zero-one loss, and, in our experiments, we choose $\tilde{a}$ according to logistic regression. We interpret $\Pr\Big(\hat{Y}{=}1\Big)$ as a new, scalar feature value $X \in \mathbf{R}$ mapped from from higher-dimensional features X' as the output of a learned "preprocessing" function $\tilde{a}$, Assumption A.1 is as hard to satisfy in general as solving the Bayes-optimal binary classification task over higher-dimensional features. Nonetheless, we expect Assumption A.1 to be approximately satisfied by such a "preprocessing" pipeline.

A.5 Linearity of Dynamics

`L-UCBFair` relies on Assumption C.3, which asserts the existence of some Hilbert space in which the state dynamics $\mathbf{P}$ are linear. Such linearity for real-world (continuous time) dynamics holds only in infinite-dimensional Hilbert space (Brunton et al., 2021) and is not computationally tractable. In addition, the "feature map" ϕ that maps state-action pairs to the aforementioned Hilbert space must be learned by the policy maker. In experiment, we use a neural network to estimate a feature map $\hat{\phi}$ which approximately satisfies the linear MDP assumption. We defer details to Appendix F.1.

B Experimental Results

Do RL agents learn to seek favorable equilibria against short-term utility? Is a Lagrangian relaxation of long-term fairness sufficient to encourage this behavior? We give positive demonstrations for both questions.

B.1 Losses and Disparities Considered

Our experiments consider losses $\mathscr{L}$ which combine true-positive and true-negative rates, where, for $\alpha, \beta \in [0, 1]$,

$$\mathscr{L}(s, a) = 1 - \alpha\mathtt{tp}(s, a) + \beta\mathtt{tn}(s, a), \tag{4}$$

where $\mathtt{tp}(s, a) = \Pr_{s,a}(\hat{Y}{=}1, Y{=}1)$ and $\mathtt{tn}(s, a) = \Pr_{s,a}(\hat{Y}{=}{-1}, Y{=}{-1})$. For disparity $\mathscr{D}$, we consider demographic parity (`DP`) (Dwork et al., 2012), equal opportunity (`EOp`) (Hardt et al., 2016), and qualification rate (`QR`):
`QR` does not matter to myopic fair classification, which does not consider mutable population state.

Func.	Form	$\xi^y_{s,a}(g) = \Pr_{s,a}(\cdot)$
DP		$\hat{Y}{=}1 \mid G{=}g$
QR	$\left\|\xi_{s,a}(g_1){-}\xi_{s,a}(g_2)\right\|^2/2$	$\hat{Y}{=}1 \mid G{=}g$
EOp		$\hat{Y}{=}1 \mid G{=}g$
EO	$\sum_y \left\|\xi^y_{s,a}(g_1) - \xi^y_{s,a}(g_2)\right\|^2/2$	$\hat{Y}{=}\hat{y} \mid Y{=}y, G{=}g$

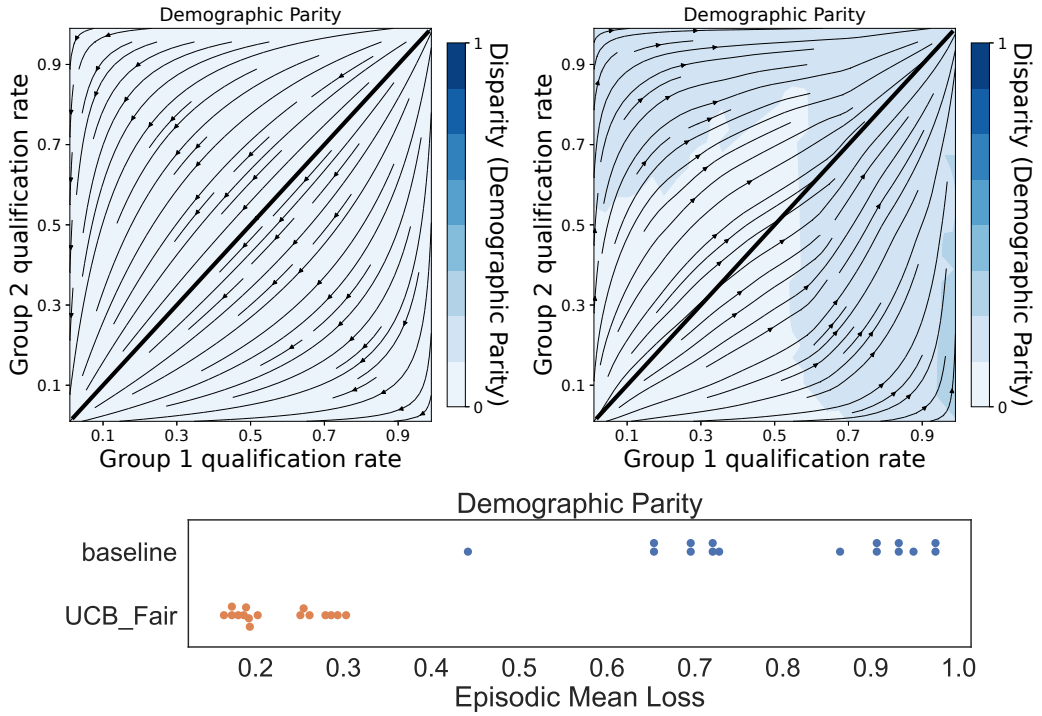


Figure 2: The greedy baseline algorithm (left) and L-UCBFair (right) are tasked to maximize the fraction of true-positive classifications ($\mathscr{L} = 1 - \texttt{tp}$, Eq. (4)), subject to demographic parity ($\mathscr{D}{=}\texttt{DP}$, Appendix B.1). The greedy algorithm uses $\lambda{=}0.5$ in Eq. (3), while L-UCBFair is trained for 2,000 steps on episodes of length 100 prior to generating this "phase portrait". We depict the expected dynamics (averaged over 20 policy iterations for each state) of the classifier-population system, parameterized by the time-evolving qualification rate in each group (1 on the horizontal, 2 on the vertical). Each group is of equal size and identically modeled by the standard normal $X \sim \mathcal{N}(Y, 1)$. Note that states in the left plot attract to universal non-qualification $\Pr(Y{=}1){=}0$, while the right plot converges to universal qualification. The lower plot shows average loss over pairs of randomly sampled episodes.

B.2 Results

Our experiments show that algorithms trained with an RL formulation of long-term fairness can drive a reactive population toward states with higher utility and fairness, even when short-term utility is *misaligned* with desirable dynamics. Our central hypothesis, that long-term fairness via RL may induce an algorithm to sacrifice short-term utility for better long-term outcomes, is concretely demonstrated by Fig. 2, in which a greedy classifier and L-UCBFair, maximizing true positive rate tp (Appendix B.1) subject to demographic parity DP (Appendix B.1), drive a population to universal non-qualification ($\Pr(Y{=}1) \to 0$) and universal qualification ($\Pr(Y{=}1) \to 1$), respectively. Each phase plot shows the dynamics of qualification rates $q_g = \Pr(Y{=}1 \mid G{=}g)$, which parameterize the population state s and define the axes, with streamlines; color depicts averaged disparity $\mathscr{D}$ incurred by a in state s.

While both algorithms achieve no or little violation of demographic parity, the myopic algorithm eventually precludes future true-positive classifications (arrows in Fig. 2 approach a low qualification state), while L-UCBFair maintains stochastic thresholds at equilibrium (mean [0.49, 0.38], by

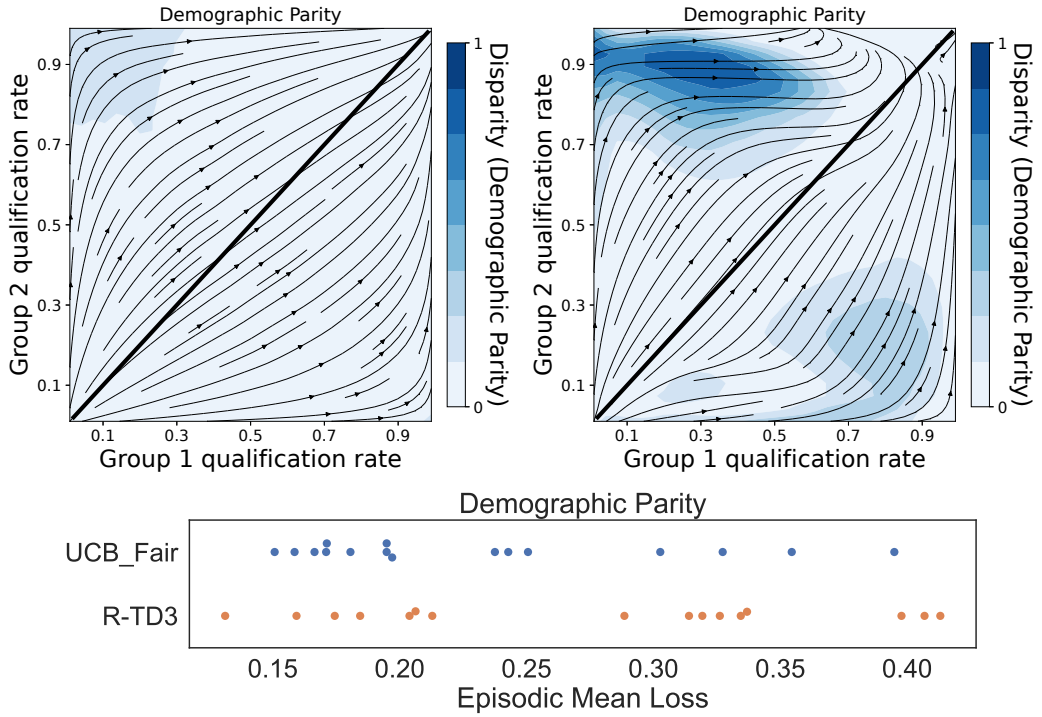


Figure 3: Using a modelled population initialized with the "Adult" dataset, reweighted for equal group representation (Appendix A.4), `L-UCBFair` (left) and `R-TD3` (right) are tasked, as in Fig. 2, to maximize the fraction of true-positive classifications ($\mathscr{L} = 1 - \texttt{tp}$, Eq. (4)), subject to demographic parity ($\mathscr{D}$=`DP`, Appendix B.1). `L-UCBFair` performs almost indistinguishably from the experiment on the synthetic dataset (Fig. 2), while `R-TD3` learns qualitatively similar behavior with more aggressive short-term violations of the fairness constraint.

group) with a non-trivial fraction of true-positives. The episodic mean loss and disparity training curves for `L-UCBFair` are depicted in Fig. 4.

We show that RL algorithms that are not limited by the same restrictive assumptions as `L-UCBFair` are applicable to long-term fairness. In Fig. 3, `R-TD3` achieves similar qualitative behavior (i.e., driving near-universal qualification at the expense of short-term utility) when optimizing a loss subject to scheduled disparity regularization. This figure also highlights the lack of guarantees of `R-TD3` in incurring prominent violations of the fairness constraint and failing to convincingly asymptote to the global optimum.

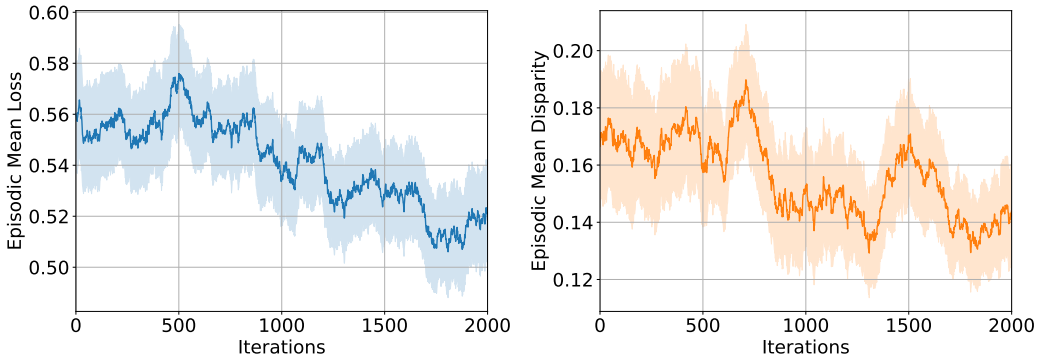


Figure 4: `L-UCBFair` 20-step sliding mean & std training loss (left) and disparity (right) for the Fig. 2 setting.

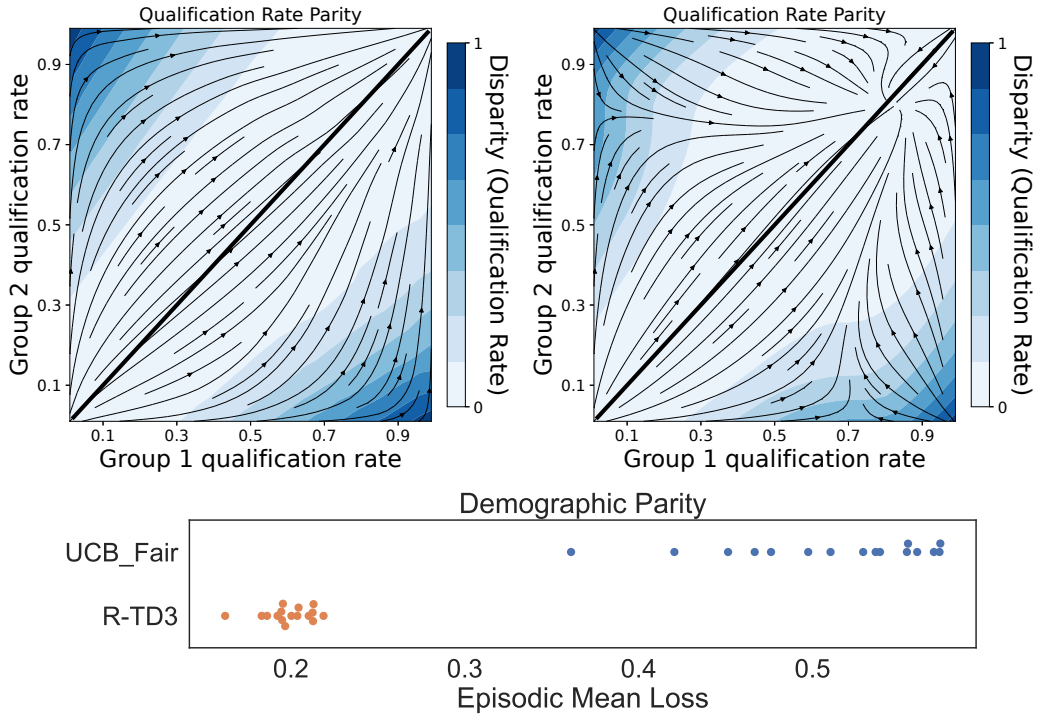


Figure 5: Phase portraits for `L-UCBFair` (left), and `R-TD3` (right) interacting on the synthetic distribution $X \sim \mathcal{N}(Y, 1)$ with groups of equal size. Both algorithms use $\mathscr{L} = 1 - \texttt{tp} - \texttt{tn}$ (i.e., zero-one loss) and $\mathscr{D} = \texttt{QR}$. Shading: qualification rate disparity for the *next* time-step.

Finally, we demonstrate the capability of RL to utilize notions of fairness that are impossible to treat in the myopic setting, viz. qualification rate parity, in Fig. 5. In this example, while both RL agents achieve qualification rate parity, we note that `L-UCBFair` fails to realize the optimal equilibrium discovered by `R-TD3`.

C TECHNICAL ASSUMPTIONS

Assumption C.1 (Slater's Condition)**.** $\exists\, \gamma > 0$, $\bar{\pi}$, such that $V_{g,1}^{\bar{\pi}}(s) \geq \tilde{c} + \gamma$.

Slater's condition is also adopted by Efroni et al. (2020); Ding et al. (2021); Ghosh et al. (2022).

Assumption C.2 (Boundedness of ν^*)**.** For $\bar{\pi}$ and $\gamma > 0$ satisfying Slater's Condition (Assumption C.1), $\nu^* \leq \frac{V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi}(s_1)}{\gamma} \leq \frac{H}{\gamma} \coloneqq \mathscr{V}$.

Assumption C.2 defines $H/\gamma = \mathscr{V}$ as an upper bound for the optimal dual variable ν^*. $\mathscr{V}$ is an input to `L-UCBFair`.

Assumption C.3 (Linear MDP)**.** MDP$(\mathcal{S}, \mathcal{A}, H, \mathbf{P}, \mathscr{L}, \mathscr{D})$ is a linear MDP with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$: For any h, there exist d signed measures $\mu_h = \{\mu_h^1, \ldots, \mu_h^d\}$ over $\mathcal{S}$, such that, for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \mu_h(s') \rangle$, In addition, there exist vectors $\theta_{r,h}, \theta_{g,h} \in \mathbb{R}^d$, such that, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, $r(s, a) = \langle \phi(s, a), \theta_{r,h} \rangle$; $\quad g(s, a) = \langle \phi(s, a), \theta_{g,h} \rangle$

Assumption C.3 addresses the curse of dimensionality when state space $\mathcal{S}$ is the space of distributions over X, Y, G. This assumption is also used in (Jin et al., 2020; Ghosh et al., 2022), with a similar assumption made in (Ding et al., 2021).

D PROOFS

Without loss of generality, we assume $\|\phi(s, a)\| \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ for all $h \in [H]$.

Lemma D.1. *The Voronoi partitioning described above satisfies $\mathcal{I}_i \cap \mathcal{I}_j = \varnothing, \forall i \neq j$ and $\cup_{i=1}^{M} \mathcal{I}_i = \mathcal{A}$. Additionally, if the number M of distinct loci or regions partitioning $\mathcal{A}$ is sufficiently large, there exists a set of loci I such that $\forall a \in \mathcal{I}_i, i \in M, \|a - I_i\|_2 \leq \epsilon_I$.*

D.1 PROOF OF THEOREM 3.3

Theorem 3.3 (Boundedness)**.** *With probability $1 - p$, there exists a constant b such that `L-UCBFair` (Algorithm 1) achieves Regret$(K) = \tilde{\mathcal{O}}\left(H^2\sqrt{d^3K}\right)$, Distortion$(K) = \tilde{\mathcal{O}}\left(H^2\sqrt{d^3K}\right)$.*

Outline The outline of this proof simulates the proof in Ghosh et al. (2022). For brevity, denote $\mathbb{P}_h V_{j,h+1}^{\pi}(s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot \mid s, a)} V_{j,h+1}^{\pi}(s')$ for $j = r, g$. Then

$$Q_{j,h}^{\pi}(s, a) = \left(r_h + \mathbb{P}_h V_{j,h+1}^{\pi}\right)(s, a) \tag{5}$$

$$V_{j,h}^{\pi}(s) = \left\langle \pi_h(\cdot \mid s) Q_{j,h}^{\pi}(s, \cdot) \right\rangle_{\mathcal{A}} \tag{6}$$

$$\left\langle \pi_h(\cdot \mid s), Q_{j,h}^{\pi}(s, \cdot) \right\rangle_{\mathcal{A}} = \sum_{a \in \mathcal{A}} \pi_h(a \mid s) Q_{j,h}^{\pi}(s, a) \tag{7}$$

Similar to Efroni et al. (2020), we establish

$$\begin{aligned}
&\text{Regret}(K) + \nu\text{Distortion}(K) \\
=&\sum_{k=1}^{K}\left(V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi_k}(s_1)\right) + \nu\sum_{k=1}^{K}\left(b - V_{g,1}^{\pi_k}(s_1)\right) \\
\leq&\underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{\pi^*}(s_1) + \nu_k V_{g,1}^{\pi^*}(s_1)\right) - \left(V_{r,1}^{k}(s_1) + \nu_k V_{g,1}^{k}(s_1)\right)}_{\mathcal{T}_1} \\
&+\underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi_k}(s_1)\right) + \nu\sum_{k=1}^{K}\left(V_{g,1}^{k}(s_1) - V_{g,1}^{\pi_k}(s_1)\right)}_{\mathcal{T}_2} \\
&+\underbrace{\frac{1}{2\eta}\nu^2 + \frac{\eta}{2}H^2K}_{\mathcal{T}_3}
\end{aligned} \tag{8}$$

$\mathcal{T}_3$ is easily bounded if η. The major task remains bound $\mathcal{T}_1$ and $\mathcal{T}_2$.

Bound $\mathcal{T}_1$ and $\mathcal{T}_2$. We have following two lemmas.

Lemma D.2 (Boundedness of $\mathcal{T}_1$)**.** *With probability $1 - p/2$, we have $\mathcal{T}_1 \leq KH\left(\frac{\log(M)}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dK}{\varsigma}}\right)$. Specifically, if $\alpha = \frac{\log(M)K}{2(1+\mathscr{V}+H)}$ and $\varsigma = 1$, we have $\mathcal{T}_1 \leq 2H(1+\mathscr{V}+H) + 2KH^2(1+\mathscr{V})\rho\epsilon_I\sqrt{dK}$ with probability $1 - p/2$.*

Lemma D.3 (Boundedness of $\mathcal{T}_2$)**.** *(Ghosh et al., 2022) With probability $1 - p/2, \mathcal{T}_2 \leq \mathcal{O}\left((\nu+1)H^2\zeta\sqrt{d^3K}\right)$, where $\zeta = \log[\log(M)4dHK/p]$.*

Lemma D.3 follows the same logic in Ghosh et al. (2022), and we delay the proof of Lemma D.2 to Appendix D.3. Now we are ready to proof Theorem 3.3.

Proof. For any $\nu \in [0, \mathscr{V}]$, with prob. $1 - p$,

$$\begin{aligned}
&\text{Regret}(K) + \nu\text{Distortion}(K) \\
\leq&\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 \\
\leq&\frac{1}{2\eta}\nu^2 + \frac{\eta}{2}H^2K + \frac{HK\log M}{\alpha} + 2KH^2(1+\mathscr{V})\rho\epsilon_I\sqrt{dK} + \mathcal{O}\left((\nu+1)H^2\zeta\sqrt{d^3K}\right)
\end{aligned} \tag{9}$$

Taking $\nu = 0, \eta = \frac{\mathscr{V}}{\sqrt{KH^2}}, \alpha = \frac{K\log M}{2(1+\mathscr{V}+H)}, \epsilon_I = \frac{1}{2\rho(1+\mathscr{V})KH\sqrt{d}}$, there exist constant b,

$$\begin{aligned}
\text{Regret}(K) \leq& \frac{\mathscr{V}H}{2}\sqrt{K} + 2H(1+\mathscr{V}+H) + 2H^2K(1+\mathscr{V})\rho\epsilon_I\sqrt{dK} + \mathcal{O}\left(H^2\zeta\sqrt{d^3K}\right) \\
\leq& \left(b\zeta H^2\sqrt{d^3} + (\mathscr{V}+1)H\right)\sqrt{K} = \tilde{\mathcal{O}}(H^2\sqrt{d^3K}).
\end{aligned}$$

Taking $\nu = \mathscr{V}, \eta = \frac{\mathscr{V}}{\sqrt{KH^2}}, \alpha = \frac{K\log M}{2(1+\mathscr{V}+H)}, \epsilon_I = \frac{1}{2\rho(1+\mathscr{V})KH\sqrt{d}}$,

$$\text{Regret}(K) + \mathscr{V}\text{Distortion}(K) \leq (\mathscr{V}+1)H\sqrt{K} + (1+\mathscr{V})\mathcal{O}\left(H^2\zeta\sqrt{d^3K}\right)$$

Following the idea of Efroni et al. (2020), there exists a policy π' such that $V_{r,1}^{\pi'} = \frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi_k}, V_{g,1}^{\pi'} = \frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi_k}$. By the occupancy measure, $V_{r,1}^{\pi}$ and $V_{g,1}^{\pi}$ are linear in occupancy measure induced by π. Thus, the average of K occupancy measure also produces an occupancy measure which induces policy π' and $V_{r,1}^{\pi'}$, and $V_{g,1}^{\pi'}$. We take $\nu = 0$ when $\sum_{k=1}^{K}\left(b - V_{g,1}^{\pi_k}\left(s_1^k\right)\right) < 0$,

otherwise $\nu = \mathscr{V}$. Hence, we have

$$\begin{aligned} & V_{r,1}^{\pi^*}(s_1) - \frac{1}{K}\sum_{k=1}^{K} V_{r,1}^{\pi_k}(s_1) + \mathscr{V}\max\left((c - \frac{1}{K}\sum_{k=1}^{K} V_{g,1}^{\pi_k}(s_1), 0\right) \\ & = V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi'}(s_1) + \mathscr{V}\max\left(c - V_{g,1}^{\pi'}(s_1), 0\right) \\ & \leq \frac{\mathscr{V}+1}{K} H\sqrt{K} + \frac{\mathscr{V}+1}{K}\mathcal{O}\left(H^2\zeta\sqrt{d^3K}\right) \end{aligned} \tag{10}$$

Since $\mathscr{V} = 2H/\gamma$, and using the result of Lemma D.12, we have

$$\max\left(c - \frac{1}{K}\sum_{k=1}^{K} V_{g,1}^{\pi_k}(s_1^k), 0\right) \leq \frac{\mathscr{V}+1}{K\mathscr{V}}\mathcal{O}\left(H^2\zeta\sqrt{d^3K}\right)$$

In this section we proof Lemma D.2 and Lemma D.3.

D.2 Prepare for Lemma D.2

In order to bound $\mathcal{T}_1$ and $\mathcal{T}_2$, we introduce the following lemma.

Lemma D.4. *There exists a constant B_2 such that for any fixed $p \in (0,1)$, with probability at least $1 - p/2$, the following event holds*

$$\|\sum_{\tau=1}^{k-1}\phi_{j,h}^{\tau}\left[V_{j,h+1}^{k}(s_{h+1}^{\tau}) - \mathbb{P}_h V_{j,h+1}^{k}(s_h^{\tau}, a_h^{\tau})\right]\|_{(\varsigma_h^k)^{-1}} \leq B_2 dHq$$

for $j \in \{r, g\}$, where $q = \sqrt{\log\left[4(B_1+1)\log(M)dT/p\right]}$ for some constant B_1.

We delay the proof of Lemma D.4 to Appendix D.4.

Lemma D.4 shows the bound of estimated value function $V_{j,h}^k$ and value function $V_{j,h}^{\pi}$ corresponding in a given policy at k. We now introcuce the following lemma appeared in Ghosh et al. (2022). This lemma bounds the difference between the value function without bonus in `L-UCBFair` and the true value function of any policy π. This is bounded using their expected difference at next step, plus a error term.

Lemma D.5. *(Ghosh et al., 2022) There exists an absolute constant $\beta = C_1 dH\sqrt{\zeta}, \zeta = \log(\log(M)4dT/p)$, and for any fixed policy π, for the event defined in Lemma D.4, we have*

$$\left\langle \phi(s,a), w_{j,h}^k \right\rangle - Q_{j,h}^{\pi}(s,a) = \mathbb{P}_h\left(V_{j,h+1}^k - V_{j,h+1}^{\pi}\right)(s,a) + \Delta_h^k(s,a)$$

for some $\Delta_h^k(s,a)$ that satisfies $\left|\Delta_h^k(s,a)\right| \leq \beta\sqrt{\phi(s,a)^T\left(\Lambda_h^k\right)^{-1}\phi(s,a)}$.

Lemma D.6. *(Ghosh et al., 2022) With probability at least $1 - p/2$, (for the event defined in Lemma D.4)*

$$Q_{r,h}^{\pi}(s,a) + \nu_k Q_{g,h}^{\pi}(s,a) \leq Q_{r,h}^{k}(s,a) + \nu_k Q_{g,h}^{k}(s,a) - \mathbb{P}_h\left(V_{h+1}^k - V_{h+1}^{\pi,\nu_k}\right)(s,a)$$

We also introduce the following lemma. This lemma bound the value function by taking `L-UCBFair` policy and greedy policy.

Lemma D.7. *Define $\bar{V}_h^k(\cdot) = \max_a\left[Q_{r,h}^k(\cdot,a) + \nu_k Q_{g,h}^k(\cdot,a)\right]$ the value function corresponding to greedy policy, we have*

$$\bar{V}_h^k(s) - V_h^k(s) \leq \frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}. \tag{11}$$

Proof. Define a_g the solution of greedy policy,

$$V_h^k(s) - V_h^k(s) = \left[Q_{r,h}^k(s, a_g) + \nu_k Q_{g,h}^k(s, a_g)\right] \tag{12}$$

$$- \int_a \pi_{h,k}(a \mid s) \left[Q_{r,h}^k(s,a) + \nu_k Q_{g,h}^k(s,a)\right] da \tag{13}$$

$$\leq \left[Q_{r,h}^k(s, a_g) + \nu_k Q_{g,h}^k(s, a_g)\right] \tag{14}$$

$$- \sum_i \mathrm{SM}_\alpha(I_i \mid x) \left[Q_{r,h}^k(x, I_i) + \nu_k Q_{g,h}^k(x, I_i)\right] + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}} \tag{15}$$

$$\leq \left(\frac{\log\left(\sum_a \exp\left(\alpha\left(Q_{r,h}^k(s, I_i) + \nu_k Q_{g,h}^k(s, I_i)\right)\right)\right)}{\alpha}\right) \tag{16}$$

$$- \sum_i \mathrm{SM}_\alpha(I_i \mid s) \left[Q_{r,h}^k(s, I_i) + \nu_k Q_{g,h}^k(s, I_i)\right] + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}} \tag{17}$$

$$\leq \frac{\log(M)}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}. \tag{18}$$

The first inequality follows from Lemma D.11 and the second inequality holds because of Proposition 1 in Pan et al. (2019).

D.3 Proof of Lemma D.2

Now we're ready to proof Lemma D.2.

Proof. This proof simulates Lemma 3 in Ghosh et al. (2022).

We use induction to proof this lemma. At step H, we have $Q_{j,H+1}^k = 0 = Q_{j,H+1}^\pi$ by definition. Under the event in Lemma D.10 and using Lemma D.5, we have for $j = r, g$,

$$\left|\left\langle \phi(s,a), w_{j,H}^k(s,a)\right\rangle - Q_{j,H}^\pi(s,a)\right| \leq \beta\sqrt{\phi(s,a)^T\left(\Lambda_H^k\right)^{-1}\phi(s,a)}$$

Thus $Q_{j,H}^\pi(s,a) \leq \min\left\{\left\langle \phi(s,a), w_{j,H}^k\right\rangle + \beta\sqrt{\phi(s,a)^T\left(\Lambda_H^k\right)^{-1}\phi(s,a)}, H\right\} = Q_{j,H}^k(s,a)$.

From the definition of $\bar{V}_h^k$,

$$\bar{V}_H^k(s) = \max_a \left[Q_{r,H}^k(s,a) + \nu_k Q_{g,h}^k(s,a)\right] \geq \sum_a \pi(a \mid x)\left[Q_{r,H}^\pi(s,a) + \nu_k Q_{g,H}^\pi(s,a)\right] = V_H^{\pi,\nu_k}(s)$$

for any policy π. Thus, it also holds for π^*, the optimal policy. Using Lemma D.7 we can get

$$V_H^{\pi^*,\nu_k}(s) - V_H^k(s) \leq \frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}$$

Now, suppose that it is true till the step $h+1$ and consider the step h. Since, it is true till step $h+1$, thus, for any policy π,

$$\mathbb{P}_h\left(V_{h+1}^{\pi,\nu_k} - V_{h+1}^k\right)(s,a) \leq (H-h)\left(\frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}\right)$$

From (27) in Lemma 10 and the above result, we have for any (s,a)

$$Q_{r,h}^\pi(s,a) + \nu_k Q_{g,h}^\pi(s,a) \leq Q_{r,h}^k(s,a) + \nu_k Q_{g,h}^k(s,a) + (H-h)\left(\frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}\right)$$

Hence,

$$V_h^{\pi,\nu_k}(s) \leq \bar{V}_h^k(s) + (H-h)\left(\frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}\right)$$

Now, again from Lemma 11, we have $\bar{V}_h^k(s) - V_h^k(s) \leq \frac{\log(|\mathcal{A}|)}{\alpha}$. Thus,

$$V_h^{\pi,\nu_k}(s) - V_h^k(s) \leq (H-h+1)\Big(\frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}\Big)$$

Now, since it is true for any policy π, it will be true for π^*. From the definition of V^{π,ν_k}, we have

$$\Big(V_{r,h}^{\pi^*}(s) + \nu_k V_{g,h}^{\pi^*}(s)\Big) - \big(V_{r,h}^k(s) + \nu_k V_{g,h}^k(s)\big) \leq (H-h+1)\Big(\frac{\log M}{\alpha} + 2(1+\mathscr{V})H\rho\epsilon_I\sqrt{\frac{dk}{\varsigma}}\Big)$$

Hence, the result follows by summing over K and considering $h = 1$.

D.4 Proof of Lemma D.4

We first define some useful sets. Let $\mathcal{Q}_j = \Big\{Q \mid Q(\cdot,a) = \min\Big\{w_j^T\phi(\cdot,a) + \beta\sqrt{\phi^T(\cdot,a)^T\Lambda^{-1}\phi(\cdot,a)}, H\Big\},\ a \in \mathcal{A}\Big\}$ be the set of Q functions, where $j \in \{r,g\}$. Since the minimum eigen value of Λ is no smaller than one so the Frobenius norm of Λ^{-1} is bounded.

Let $\mathcal{V}_j = \big\{V_j \mid V_j(\cdot) = \int_a \pi(a \mid \cdot)Q_j(\cdot,a)da; Q_r \in \mathcal{Q}_r, Q_g \in \mathcal{Q}_g, \nu \in [0,\mathscr{V}]\big\}$ be the set of Q functions, where $j \in \{r,g\}$. Define

$$\Pi = \left\{\pi \mid \forall a \in \mathcal{A}, \pi(a \mid \cdot) = \frac{1}{\int_{b\in\mathcal{I}(a)} db}\mathrm{SM}_\alpha\left(Q_r(\cdot,I(a)) + \nu Q_g(\cdot,I(a))\right),\ Q_r \in \mathcal{Q}_r, Q_g \in \mathcal{Q}_g, \nu \in [0,\mathscr{V}]\right\}$$

the set of policies.

It's easy to verify $V_j^k \in \mathcal{V}_j$.

Then we introduce the proof of Lemma D.4. To proof Lemma D.4, we need the ϵ-covering number for the set of value functions(Lemma D.10(Ghosh et al., 2022)). To achieve this, we need to show if two Q functions and the dual variable ν are close, then the bound of policy and value function can be derived(Lemma D.8, Lemma D.9). Though the proof of Lemma D.8 and Lemma D.9 are different from Ghosh et al. (2022), we show the results remain the same, thus Lemma D.10 still holds. We'll only introduce Lemma D.10 and omit the proof.

We now proof Lemma D.8.

Lemma D.8. *Let π be the policy of `L-UCBFair` corresponding to $Q_r^k + \nu_k Q_g^k$, i.e.,*

$$\pi(a \mid \cdot) = \frac{1}{\int_{b\in\mathcal{I}(a)} db} SM_\alpha\left(Q_r(\cdot,I(a)) + \nu Q_g(\cdot,I(a))\right) \tag{19}$$

and

$$\tilde{\pi}(a \mid \cdot) = \frac{1}{\int_{b\in\mathcal{I}(a)} db} SM_\alpha\left(\tilde{Q}_r(\cdot,I(a)) + \tilde{\nu}\tilde{Q}_g(\cdot,I(a))\right), \tag{20}$$

if $\left|Q_j - \tilde{Q}_j\right| \leq \epsilon'$ and $|\nu - \tilde{\nu}| \leq \epsilon'$, then $\left|\int_a (\pi(a \mid x) - \tilde{\pi}(a \mid x))\, da\right| \leq 2\alpha\epsilon'(1 + \mathscr{V} + H)$.

Proof.

$$\begin{aligned}
&\left|\int_a (\pi(a \mid x) - \tilde{\pi}(a \mid x))\, da\right| &(21)\\
=&\left|\sum_{i=1}^M \int_{a\in\mathcal{I}_i} (\pi(I(a) \mid x) - \tilde{\pi}(I(a) \mid x))\, da\right|\\
=&\left|\sum_{i=1}^M \int_{b\in\mathcal{I}_i} db\,(\pi(I_i \mid x) - \tilde{\pi}(I_i \mid x))\right|\\
\leq&\sum_{i=1}^M \left|SM_\alpha\big(Q_r(s,I_i) + \nu Q_g(s,I_i)\big) - SM_\alpha\big(\tilde{Q}_r(s,I_i) + \tilde{\nu}\tilde{Q}_g(s,I_i)\big)\right|\\
\leq&2\alpha\left|Q_r(\cdot,I(a)) + \nu Q_g(\cdot,I(a)) - \tilde{Q}_r(\cdot,I(a)) - \tilde{\nu}\tilde{Q}_g(\cdot,I(a))\right| &(22)
\end{aligned}$$

The last inequaity holds because of Theorem 4.4 in Epasto et al. (2020). Using Corollary D.14, we have

$$\left|\int_a (\pi(a \mid x) - \tilde{\pi}(a \mid x))\, da\right| \leq 2\alpha\epsilon'(1 + \mathscr{V} + H) \tag{23}$$

Now since we have Lemma D.8, we can further bound the value functions.

Lemma D.9. *If* $\left|\tilde{Q}_j - Q_j^k\right| \leq \epsilon'$, *where* $\tilde{Q}_j \in \mathcal{Q}_j$, *then there exists* $\tilde{V}_j \in \mathcal{V}_j$ *such that*

$$\left|V_j^k - \widetilde{V}_j\right| \leq H2\alpha\epsilon'(1 + \mathscr{V} + H) + \epsilon',$$

Proof. For any x,

$$\begin{aligned}
&V_j^k(s) - \widetilde{V}_j(s) \\
&= \left|\int_a \pi(a \mid s) Q_j^k(s,a) da - \int_a \tilde{\pi}(a \mid s) \tilde{Q}_j(s,a) da\right| \\
&= \left|\int_a \pi(a \mid s) Q_j^k(s,a) da - \int_a \pi(a \mid s) \tilde{Q}_j(s,a) da + \int_a \pi(a \mid s) \tilde{Q}_j(s,a) da - \int_a \tilde{\pi}(a \mid s) \tilde{Q}_j(s,a) da\right| \\
&\leq \left|\int_a \pi(a \mid s) \left(Q_j^k(s,a) - \tilde{Q}_j(s,a)\right) da\right| + \left|\int_a \pi(a \mid s) \tilde{Q}_j(s,a) da - \int_a \tilde{\pi}(a \mid s) \tilde{Q}_j(s,a) da\right| \\
&\leq \epsilon' + H \left|\int_a (\pi(a \mid s) - \tilde{\pi}(a \mid s))\, da\right| \\
&\leq \epsilon' + H2\alpha\epsilon'(1 + \mathscr{V} + H)
\end{aligned}$$

Using Lemmas above, we can have the same result presented in Lemma 13 of Ghosh et al. (2022) as following.

Lemma D.10. *(Ghosh et al., 2022) There exists a* $\tilde{V}_j \in \mathcal{V}_j$ *parameterized by* $\left(\tilde{w}_r, \tilde{w}_g, \tilde{\beta}, \Lambda, \tilde{\mathscr{V}}\right)$ *such that dist* $\left(V_j, \tilde{V}_j\right) \leq \epsilon$ *where*

$$\left|V_j - \tilde{V}_j\right| = \sup_x \left|V_j(s) - \tilde{V}_r(s)\right|.$$

Let $N_\epsilon^{V_j}$ *be the* ϵ*-covering number for the set* $\mathcal{V}_j$, *then,*

$$\log N_\epsilon^{V_j} \leq d \log\left(1 + 8H\frac{\sqrt{dk}}{\sqrt{\varsigma}\epsilon'}\right) + d^2 \log\left[1 + 8d^{1/2}\beta^2 / \left(\varsigma\left(\epsilon'\right)^2\right)\right] + \log\left(1 + \frac{\mathscr{V}}{\epsilon'}\right)$$

where $\epsilon' = \frac{\epsilon}{H2\alpha(1+\mathscr{V}+H)+1}$

Lemma D.11. $|Q_{j,h}^k(s,a) - Q_{j,h}^k(s, I(a))| \leq 2H\rho\epsilon_I\sqrt{\frac{dK}{\varsigma}}$.

Proof.

$$\begin{aligned}
&\left|Q_{j,h}^k(s,a) - Q_{j,h}^k(s, I(a))\right| && (24)\\
=&\left|w_{j,h}^k(s,a)^T(\phi(s,a) - \phi(s, I(a)))\right| && (25)\\
\leq&||w_{j,h}^k(s,a)||_2 \|\phi(s,a) - \phi(s, I(a))||_2 && (26)\\
& && (27)
\end{aligned}$$

From Lemma D.13 and Assumption 3.2 we get the result.

D.5 PRELIMINARY RESULTS

Lemma D.12. *(Ding et al., 2021) Let ν^* be the optimal dual variable, and $C \geq 2\nu^*$, then, if*

$$V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi}(s_1) + C\left[c - V_{g,1}^{\pi}(s_1)\right]_+ \leq \delta,$$

we have

$$\left[c - V_{g,1}^{\tilde{\pi}}(x_1)\right]_+ \leq \frac{2\delta}{C}.$$

Lemma D.13. *(Jin et al., 2020) For any (k, h), the weight $w_{j,h}^k$ satisfies*

$$\left\|w_{j,h}^k\right\| \leq 2H\sqrt{dk/\varsigma}$$

Corollary D.14. *If* $\operatorname{dist}\left(Q_r, \tilde{Q}_r\right) \leq \epsilon', \operatorname{dist}\left(Q_g, \tilde{Q}_g\right) \leq \epsilon'$, *and* $|\tilde{\nu}_k - \nu_k| \leq \epsilon'$, *then,* $\operatorname{dist}\left(Q_r^k + \nu_k Q_g^k, \tilde{Q}_r + \tilde{\nu}_k \tilde{Q}_g\right) \leq \epsilon'(1 + \mathscr{V} + H)$.

E ADDITIONAL FIGURES

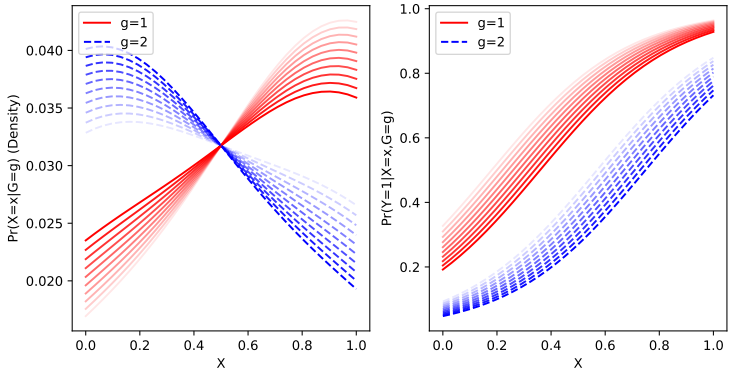


Figure 6: A synthetic distribution is updated according to a dynamical kernel $\mathbf{P}$ based on evolutionary dynamics (Appendix A.3), when a classifier repeatedly predicts $\hat{Y}{=}1$ iff $X \geq 0.5$. We visualize how the distribution of X and conditional qualification rates $\Pr(Y{=}1 \mid X)$ change in each group $g \in \{1 \text{ (red, solid)}, 2 \text{ (blue, dashed)}\}$, fading the plotted lines over 10 time steps. In this example, the feature values X in each group decrease with time, while the qualification rates of agents at any fixed value of X decrease.

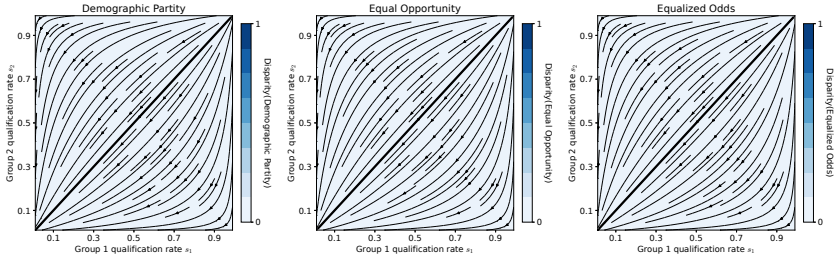
F Additional Experiment Results

F.1 Experiment Details

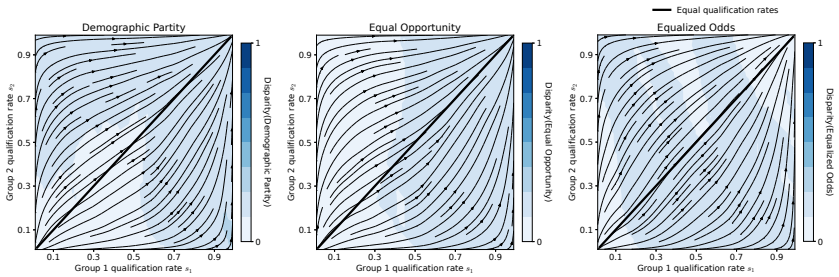
Device and Packages. We run all the experiment on a single 1080Ti GPU. We implement the `R-TD3` agent using StableBaseline3Raffin et al. (2021). The neural network is implemented using PytorchPaszke et al. (2019).

Neural Network to learn ϕ. We use a multi-layer perceptron to learn ϕ. Specifically, we sample 100000 data points using a random policy, storing s, a, r and g. The inputs of the network are state and action, passing through fully connected (fc) layers with size 256, 128, 64, 64. ReLU is used as activation function between fc layers, while a SoftMax layer is applied after the last fc layer. We treat the outcome of this network as ϕ. To learn ϕ, we apply two separated fc layers (without bias) with size 1 to $\hat{\phi}$ and treat the outputs as predicted r and predicted g. A combination of MSE losses of r and g are adopted. We use Adam as the optimizer. Weight decay is set to 1e-4 and learning rate is set to 1e-3, while batch size is 128.

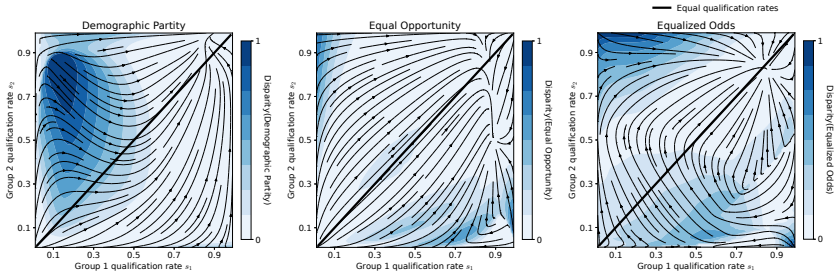
Note that, $\hat{\phi}$ is linear regarding r and g, but the linearity of transition kernel cannot be captured using such a schema. Therefore, equivalently we made an assumption that there always exists measure μ_h such that for given $\hat{\phi}$, the linearity of transition kernel holds. It's a stronger assumption than Assumption C.3.



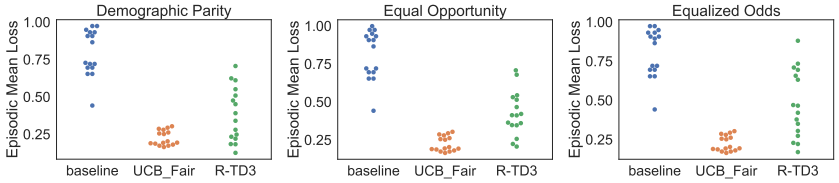
(a) A baseline, greedy classifier performing gradient descent to (locally) maximize true-positives, with static fairness regularization (columns).



(b) `L-UCBFair`, trained for 2,000 steps on the same, cumulative utility functions.

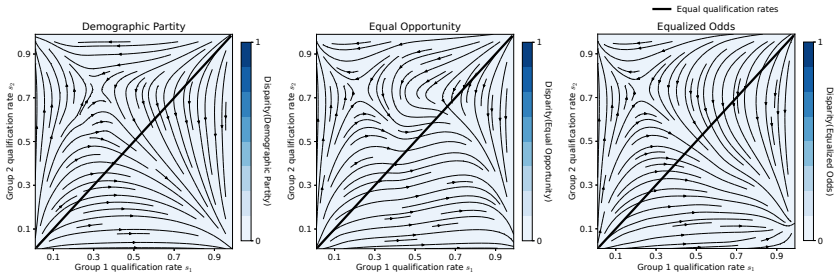


(c) A `R-TD3` agent (Section 3.2) trained for 200,000 steps on the same, cumulative utility functions.

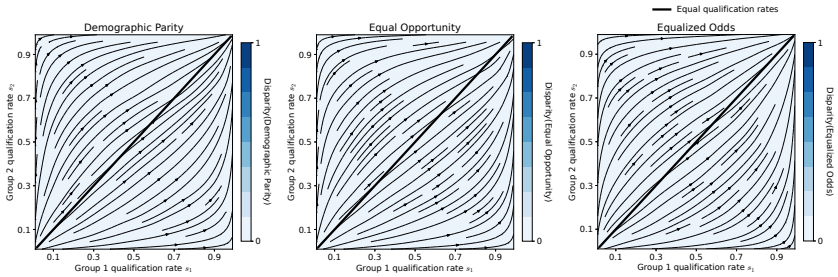


(d)

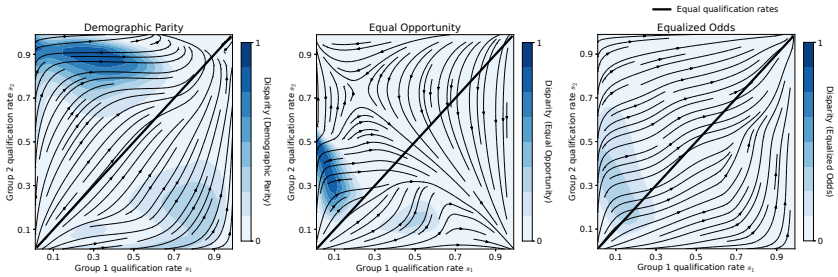
Figure 7: A comparison of learning policies trained to optimize **cumulative true positive fraction** subject to three different regularized fairness constraints (columns) with $\nu = 1$ (`L-UCBFair`), $\lambda = 0.5$, (greedy agent), and the time-dependent regularization detailed in Section 3.2 (`R-TD3`). The first policy (top row) is a baseline, myopic policy that greedily seeks to optimize current utility in any state by performing gradient decent. The second policy (bottom row) is trained using deep reinforcement learning (`R-TD3`) as detailed in Section 3.2 for 200,000 steps before we terminate learning and generate the phase portraits depicted. This is on the synthetic distribution. In all cases, the baseline, greedy policy drives the system to promote unqualified individuals, with low qualification rates in each group, while the `R-TD3` agent is able to drive the system to more favorable equilibria characterized by higher qualification rates. The shading in the phase plots depicts the violation of the regularizing fairness constraint within each column, validating the claim that the `R-TD3` agent learns to sacrifice short-term utility to drive towards preferable system states.



(a) A baseline, greedy classifier locally maximizing true positive classifications, regularized by fairness (columns).



(b) `L-UCBFair`, trained for 2,000 steps on the same, cumulative utility functions.



(c) A `R-TD3` agent (Section 3.2) trained for 200,000 steps on the same, cumulative utility functions.

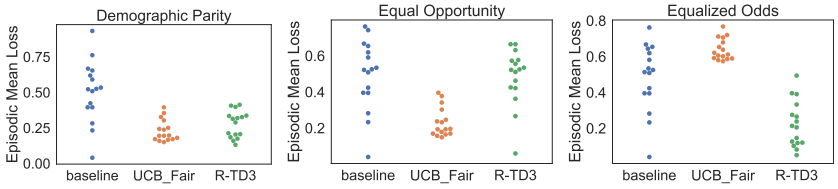
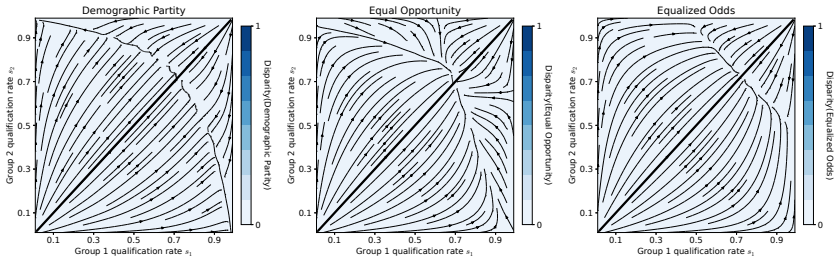
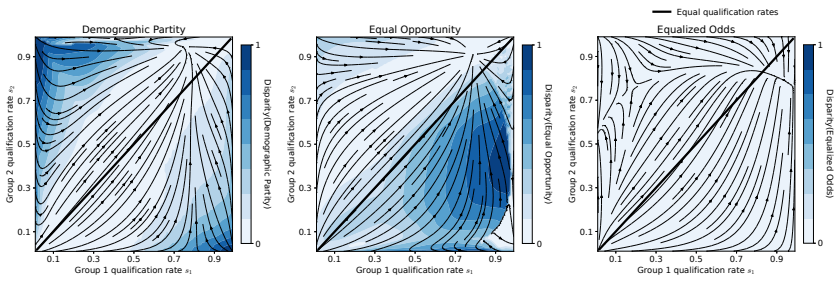


Figure 8: A repetition of the experiment performed in Fig. 7, rewarding **true positive fraction** using data synthesized from the **UCI "Adult Data Set"**, as detailed in section Appendix A.4 with equal group size reweighting. For this experiment, an individual's sex defined their "group" membership, which is an imbalanced label in the dataset ($\approx 67\%$ male, group 2, vertical axis) that we re-weight for equal representation Appendix A. The stark difference between Fig. 7 and this experiment in the qualitative behavior of the greedy agent can be largely explained by the fact that $\Pr(Y{=}1|X{=}x))$ is not actually monotonically increasing in x, as stipulated by Assumption A.1. Indeed, if $Pr(Y{=}1|X{=}x)$ is sufficiently rough, the threshold selected by the baseline agent is liable to appear as if sampled uniformly at random, which is how the initial threshold value is chosen for each of the 20 iterations averaged over for each pair of group qualification rates used to generate the phase portraits above. Despite this failure mode of the baseline agent, however, the `R-TD3` agent is still largely able to drive the system towards equilibria with more equal qualification rates in both groups. The line of equal qualification rates in both groups is depicted in black, from the lower-left corner of each phase plot to the upper-right.

F.2 ZERO ONE ACCURACY WITH MORE WEIGHTS ON TRUE POSITIVE



(a) A baseline, greedy classifier locally maximizing utility, regularized by fairness (columns).



(b) A `R-TD3` agent (Section 3.2) trained for 200,000 steps on the same, cumulative utility functions.

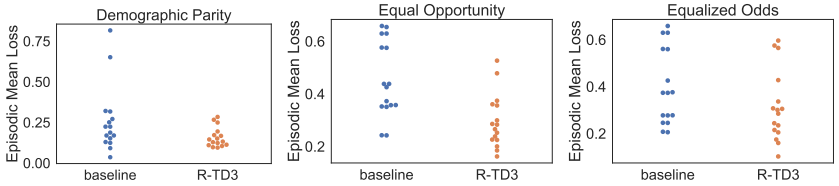
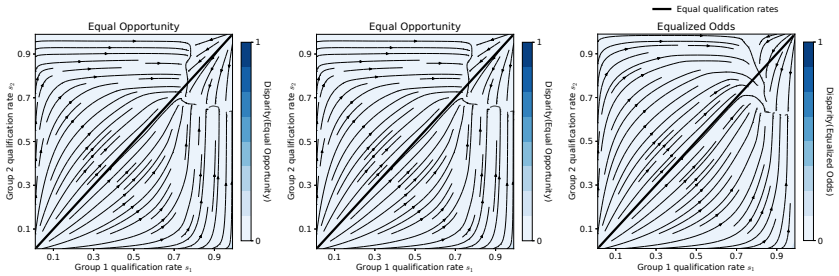
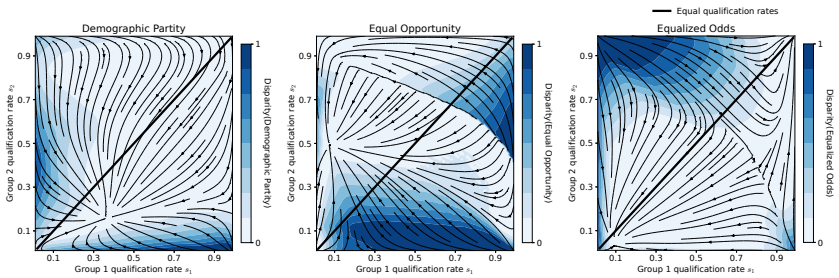


Figure 9: A repetition of the experiment performed in Fig. 7 with a different base utility function **(true positive fraction + 0.8 true negative fraction)**, weighting each regularized disparity term with $y = 1$, with the same **synthetic distribution**. While our observations are largely consistent with Fig. 7, we also note that the `R-TD3` agent drives a subset of state-space in the third pane to an equilibrium less desired than the one that the myopic agent reaches.



(a) A baseline, greedy classifier locally maximizing utility, regularized by fairness (columns).



(b) A R-TD3 agent (Section 3.2) trained for 200,000 steps on the same, cumulative utility functions.

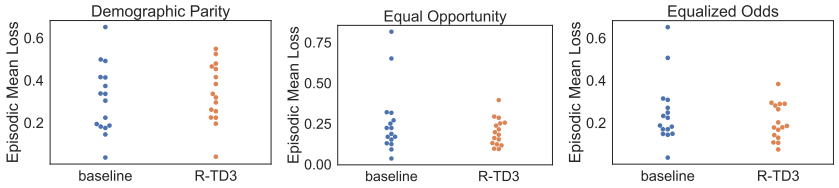
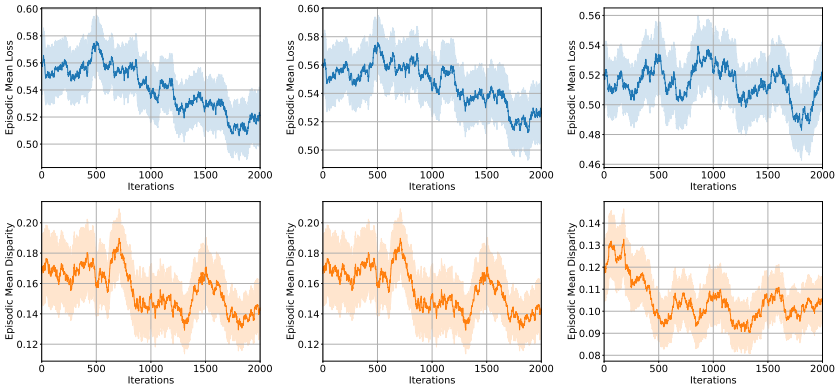


Figure 10: A repetition of the experiment performed in Fig. 9 (i.e., with a base utility function **(true positive fraction + 0.8 true negative fraction)** and $y = 1$ weighted regularized disparity term), on the **UCI "Adult Data Set"**, as detailed in section Appendix A.4 with groups re-weighted for equal representation.

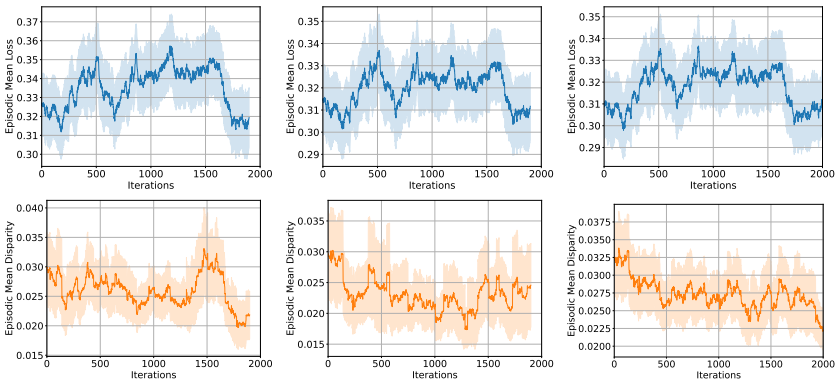
F.3 Training Curves: L-UCBFair

F.3.1



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

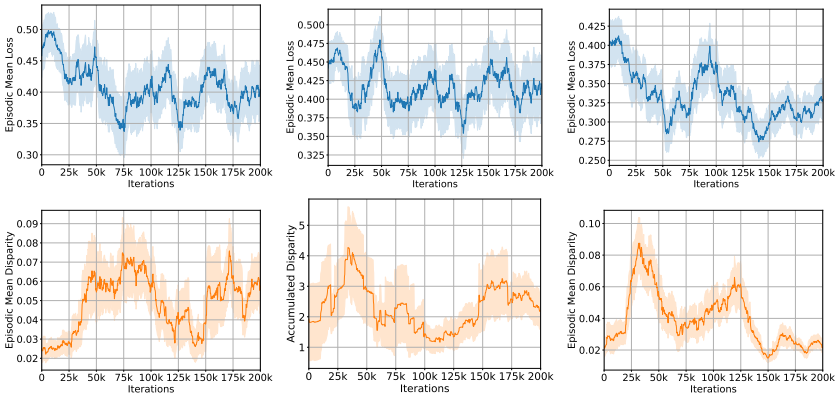
Figure 11: L-UCBFair 20-step sliding mean & std for the setting in Fig. 7.



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

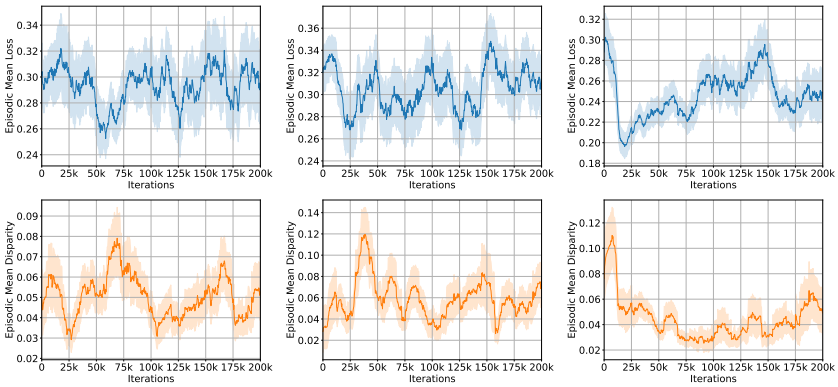
Figure 12: `L-UCBFair` 20-step sliding mean & std for the setting in Fig. 8.

F.4 Training Curves: `R-TD3`



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

Figure 13: `R-TD3` 100-step sliding mean & std for the setting in Fig. 8.



(a) Demographic Parity (b) Equal Opportunity (c) Equalized Odds

Figure 14: `R-TD3` 100-step sliding mean & std for the setting in Fig. 10.

F.5 REDUCTION OF UTILITY

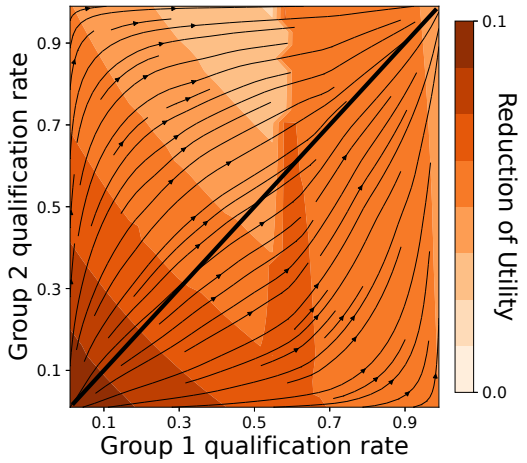


Figure 15: The figure depicts the short-term impact on utility of the `UCBFair` algorithm compared to a greedy baseline agent that operates without fairness constraints. In this experiment, both algorithms were designed to optimize the fraction of true-positive classifications, but only `UCBFair` was subject to the additional constraint of demographic parity. As the results indicate, the `UCBFair` algorithm experiences a reduction in utility compared to the greedy baseline, but it is able to drive the system towards a state that is preferable in the long term.