Algorithms for Fairness in Sequential Decision Making

Min Wen
University of Pennsylvania

Osbert Bastani
University of Pennsylvania

Ufuk Topcu
University of Texas at Austin

Abstract

It has recently been shown that if feedback effects of decisions are ignored, then imposing fairness constraints such as demographic parity or equality of opportunity can actually exacerbate unfairness. We propose to address this challenge by modeling feedback effects as Markov decision processes (MDPs). First, we propose analogs of fairness properties for the MDP setting. Second, we propose algorithms for learning fair decision-making policies for MDPs. Finally, we demonstrate the need to account for dynamical effects using simulations on a loan applicant MDP.

1 Introduction

Machine learning has the potential to substantially improve performance in tasks such as legal and financial decision-making. However, biases in the data can be reflected in a decision-making policy trained on that data (Dwork et al., 2012), which can result in decisions that unfairly discriminate against minorities. For example, consider the problem of deciding whether to give loans to applicants (Hardt et al., 2016). If minorities are historically given loans less frequently, then there may be less data on how reliably they repay loans. Thus, a learned policy may unfairly label minorities as higher risk and deny them loans.

So far, work on fairness has largely focused on supervised learning. However, naïvely imposing fairness constraints while ignoring even one-step feedback effects can actually harm minorities (Liu et al., 2018; Creager et al., 2019; D'Amour et al., 2020). Thus, we must extend existing definitions of fairness to account for the feedback effects of the decisions being made on population members. For example, denying loans to individuals may have consequences on their financial security that need to be taken into account.

This paper proposes algorithms for learning fair decision-making policies that account for feedback effects of decisions. We model these effects as the dynamics of a Markov decision process (MDP), and extend existing fairness definitions to decision-making policies for a known MDP. We distinguish the quality of outcomes for the decision-maker (e.g., the bank) from the quality of the outcomes for individuals (e.g., a loan applicant). Then, fairness properties are constraints on the average quality of outcomes for individuals in different subpopulations (e.g., majorities and minorities are offered loans at the same frequency), whereas the reward measures the quality of outcomes for the decision-maker (e.g., the bank's profit). The key challenge is that learning with a fairness constraint is much more challenging in the MDP setting due to the inherent non-convexity. Building on work on constrained MDPs (Altman, 1999; Wen and Topcu, 2018), we propose novel algorithms for learning policies that satisfy fairness constraints. In particular, we propose two algorithms. First, we propose a model-based algorithm based that has optimality guarantees, but is limited to MDPs with finite state and action spaces and satisfies a *separability* assumption saying that the sensitive attribute does not change over time. Second, we propose a model-free algorithm that is very general, but may not find the optimal policy.

We compare to two baselines that ignore dynamics: (i) an algorithm that optimistically pretends actions do not affect the state distribution (i.e., supervised learning), and (ii) an algorithm that conservatively assumes the state distribution can change adversarially on each step. In a simulation study on a loan applicant MDP based on (Hardt et al., 2016), we show that compared to our algorithm, the optimistic algorithm learns unfair policies, and the conservative algorithm learns fair but poorly performing policies. Our results demonstrate the importance of accounting for dynamics.[1]

Related work. For supervised learning, there have

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] Our code is at: https://github.com/wmgithub/fairness.

been several definitions of fairness, including demographic parity (i.e., members of the majority and minority subpopulations have equal outcomes on average) (Calders et al., 2009), equality of opportunity (i.e., *qualified* members have equal outcomes on average) (Hardt et al., 2016), individual fairness (Dwork et al., 2012), and causal fairness (i.e., protected attributes should not influence outcomes) (Kusner et al., 2017; Kilbertus et al., 2017; Nabi and Shpitser, 2018). The appropriate definition depends on the application.

There has been recent interest in fairness for sequential decision making. For instance, Liu et al. (2018) has studied one-step feedback effects, Creager et al. (2019) studies the impact of dynamics on fairness via simulations, and (D'Amour et al., 2020) proposes tools from causal inference to study fairness with dynamics. However, none of these approaches propose learning algorithms. For instance, the model in Liu et al. (2018) is highly stylized (e.g., they only consider a single time step) since their goal is to demonstrate the necessity of accounting for sequential decisions rather than study the general problem of algorithms for ensuring fairness in sequential decision-making.

In the case of unknown dynamics, there has been work in the bandit setting (Joseph et al., 2016; Hashimoto et al., 2018) and the MDP setting (Jabbari et al., 2017; Elzayn et al., 2019). However, they focus on fairness constraints for which the optimal policy is always fair, so solving for the optimal fair policy is trivial once the dynamics are known. In contrast, we are interested in the setting where fairness constraint is nontrivial even when the dynamics are known. There has been recent work studying fairness constraints (Bechavod et al., 2019; Kilbertus et al., 2019) in the setting of selective labels (Lakkaraju et al., 2017); however, there is no state in their setting. In addition, Awasthi et al. (2020) study how fairness definitions can be updated over time based on feedback; in their model, individuals do not recur across time steps as they do in ours.

There has been work on constrained MDPs (Altman, 1999; Achiam et al., 2017; Wen and Topcu, 2018). However, these approaches focus on constraints that bound some state-dependent cost function; in contrast, fairness constraints say that statistics of different groups must be equalized in some way.

2 Fairness Constraints for MDPs

Preliminaries. A *Markov decision process (MDP)* is a tuple $M = (S, A, D, P, R, \gamma)$, where $S = [n] = \{1, ..., n\}$ are the states, $A = [m]$ are the actions, $D \in \mathbb{R}^{|S|}$ is the initial state distribution (i.e., D_s is the probability of starting in state s), $P \in \mathbb{R}^{|S|\times|A|\times|S|}$ are the transitions (i.e., $P_{s,a,s'}$ is the probability of transitioning from s to s' taking action a), $R \in \mathbb{R}^{|S|\times|A|}$ are the rewards (i.e., $R_{s,a}$ is the reward obtained taking action a in state s), and $\gamma \in \mathbb{R}$ is the discount factor. Let $\pi \in \mathbb{R}^{|S|\times|A|}$ be a stochastic policy (i.e., $\pi_{s,a}$ is the probability of taking action a in state s). The induced transtions are $P^{(\pi)} \in \mathbb{R}^{|S|\times|S|}$, where $P^{(\pi)}_{s,s'} = \sum_{a\in A} \pi_{s,a} P_{s,a,s'}$. The time-discounted state distribution is

$$D^{(\pi)} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t D^{(\pi,t)}$$

where

$$D^{(\pi,t)} = \begin{cases} D & \text{if } t = 0 \\ P^{(\pi)} D^{(\pi,t-1)} & \text{otherwise,} \end{cases}$$

and the time-discounted state-action distribution is $\Lambda \in \mathbb{R}^{|S|\times|A|}$, where $\Lambda^{(\pi)}_{s,a} = D^{(\pi)}_s \pi_{s,a}$. Note that $\sum_a \pi_{s,a} = 1$ and $\sum_s D^{(\pi)}_s = 1$, so $\sum_{s,a} \Lambda_{s,a} = 1$. The cumulative expected reward is

$$R^{(\pi)} = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \langle R, \Lambda^{(\pi,t)}\rangle = \mathbb{E}_{(s,a)\sim\Lambda^{(\pi)}}[R_{s,a}],$$

where $\langle X, Y\rangle = \sum_{s\in S}\sum_{a\in A} X_{s,a} Y_{s,a}$; we include a normalizing constant of $1-\gamma$ to simplify notation, which does not affect the reinforcement learning problem since $R^{(\pi)}$ is scaled equally for different policies. Given policy class Π, the optimal policy is $\pi^* = \arg\max_{\pi\in\Pi} R^{(\pi)}$.

Fairness. Consider a population of individuals (e.g., loan applicants) interacting with a decision-maker (e.g., a bank). States S encode an individual's features (e.g., probability of repaying), actions A are interventions (e.g., loan offer), and transitions P encode state changes (e.g., changes in ability to repay). The decision-maker rewards are not always aligned with individual rewards, so we use rewards R to indicate quality of outcomes for the decision-maker (e.g., the bank's profit), and *individual rewards* $\rho \in \mathbb{R}^{|S|\times|A|}$ to indicate quality of outcomes for an individual (e.g., whether a loan is offered). The cumulative expected individual rewards is $\rho^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda^{(\pi)}}[\rho_{s,a}]$.

Our goal is to learn the optimal policy for the decision-maker under a fairness constraint on the individual rewards. In particular, we want to ensure that π does not favor the *majority subpopulation* over the *minority subpopulation*. The specific fairness constraint that should be used depends on the problem domain. We show how two constraints from the supervised learning setting can be extended to the MDP setting; as we discuss below, our results are more general.

First, we have the following extension of demographic parity to the MDP setting:

Definition 2.1. Let $\epsilon \in \mathbb{R}_+$, M be an MDP with states $S = Z \times \tilde{S}$, where $Z = \{\mathrm{maj}, \mathrm{min}\}$, and $\rho \in \mathbb{R}^{|S|\times|A|}$ be the individual rewards. For $z \in Z$, let

$$\Lambda_z^{(\pi)} = \Lambda^{(\pi)} \mid \exists \tilde{s} \in \tilde{S} \;.\; s_0 = (z, \tilde{s})$$

be the time-discounted state-action distribution conditioned on starting from an initial state s_0 in subpopulation z—i.e., s_0 has the form $s_0 = (z, \tilde{s}_0)$ for some $\tilde{s}_0 \in \tilde{S}$. More precisely,

$$(\Lambda_z^{(\pi)})_{s,a} = (D_z^{(\pi)})_s \pi_{s,a} \quad (\forall s \in S, a \in A)$$
$$(D_z^{(\pi)})_s = (1-\gamma)\sum_{t=0}^{\infty}(\gamma P^{(\pi)})^t D_z \quad (\forall s \in S)$$
$$(D_z)_{s_0} = \nu^{-1} \cdot D_{s_0} \cdot \mathbb{I}[\exists \tilde{s} \in \tilde{S} \;.\; s_0 = (z, \tilde{s})] \quad (\forall s_0 \in S),$$

where ν is a normalizing constant. Furthermore, let $\rho^{(\pi)}$ conditioned on starting in subpopulation z is $\rho_z^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}]$. Then, we say a policy π satisfies ϵ *demographic parity* if $|\rho_{\mathrm{maj}}^{(\pi)} - \rho_{\mathrm{min}}^{(\pi)}| \leq \epsilon$.

For an individual $(\tilde{s}, z) \in S$, z encodes whether they are from the majority ($z = \mathrm{maj}$) or minority ($z = \mathrm{min}$) subpopulation and $\tilde{s}$ encodes their non-sensitive characteristics (e.g., probability of repaying a loan); demographic parity says the cumulative expected individual rewards are equal for the majority and minority subpopulations. Next, we have the following analog of equal opportunity (Hardt et al., 2016):

Definition 2.2. Let $\epsilon \in \mathbb{R}_+$, let M be an MDP with states $S = Z \times Y \times \tilde{S}$, where $Z = \{\mathrm{maj}, \mathrm{min}\}$ and $Y = \{\mathrm{qual}, \mathrm{unqual}\}$, and let $\rho \in \mathbb{R}^{|S|\times|A|}$ be the individual rewards. For each $z \in Z$, let

$$\rho_z^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}]$$
$$\Lambda_z^{(\pi)} = \Lambda^{(\pi)} \mid \exists \tilde{s} \in \tilde{S} \;.\; s_0 = (z, \mathrm{qual}, \tilde{s}).$$

A policy π is ϵ *equal opportunity* if $|\rho_{\mathrm{maj}}^{(\pi)} - \rho_{\mathrm{min}}^{(\pi)}| \leq \epsilon$.

This property is similar to demographic parity, but where $\Lambda_z^{(\pi)}$ is restricted to the *qualified* subpopulation (i.e., $y = \mathrm{qual}$). In other words, this property says that cumulative expected individual rewards are equal on average for qualified members of the majority and minority subpopulations.

Remark 2.3. In general, our algorithms apply to any fairness constraint that two subpopulations should have equal expected outcomes—i.e., for any $S_{\mathrm{maj}}, S_{\mathrm{min}} \subseteq S$, letting $\rho_z^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}]$ and $\Lambda_z^{(\pi)} = \Lambda^{(\pi)} \mid \mathbb{I}[s_0 \in S_z]$, the constraint $|\rho_{\mathrm{maj}}^{(\pi)} - \rho_{\mathrm{min}}^{(\pi)}| \leq \epsilon$. They also extend to one-sided inequalities and to multiple majority and minority subpopulations. They also extend to batch decisions; see Appendix A.

We focus on demographic parity when describing our algorithms, but our results are general. Letting $\Pi_{\mathrm{DP},\epsilon}$ be the class of policies satisfying demographic parity, our goal is to compute the optimal policy

$$\pi_{\mathrm{DP}}^* = \operatorname*{arg\,max}_{\pi \in \Pi_{\mathrm{DP},\epsilon}} R^{(\pi)}. \quad (1)$$

We primarily focus is on settings where the MDP is known, which includes settings where the decision-maker learns about individuals via their interactions (see example below), but not ones where they learn across individuals. We describe a basic extension to unknown MDPs in Section 5.

Example. We describe an MDP M_{loan} that models individuals applying for loans. We assume each individual has a true probability p of repaying their loan. On step t, the bank has an estimate of the distribution of p (e.g., a credit score); we assume this distribution is a Beta distribution—i.e., $p_t \sim \mathrm{Beta}(\alpha_t, \beta_t)$. Thus, the states of our MDP (α_t, β_t). [2] The actions are to offer ($a = 1$) or deny ($a = 0$) a loan. If the bank offers a loan, the transitions are

$$(\alpha_{t+1}, \beta_{t+1}) = \begin{cases} (\alpha_t + 1, \beta_t) & \text{with probability } p_t \\ (\alpha_t, \beta_t + 1) & \text{with probability } 1 - p_t. \end{cases}$$

If the bank denies the loan, the transitions are $(\alpha_{t+1}, \beta_{t+1}) = (\alpha_t, \beta_t)$. However, since we are interested in detrimental effects of the bank's decisions, we assume this decision reduces the applicant's ability to pay for future loans—i.e., $(\alpha_{t+1}, \beta_{t+1}) = (\alpha_t, \beta_t + \tau)$, where $\tau \in \mathbb{R}_+$ is a hyperparameter. We assume the initial state distribution is $z \sim \mathrm{Bernoulli}(p_Z)$ and $(\alpha, \beta) \sim p_0(\alpha, \beta \mid z)$ for some $p_Z \in [0, 1]$ and some distribution p_0—i.e., the initial distribution over the parameters α, β depends on the whether the applicant is from the majority or minority subpopulation. Note that p_0 can additionally be conditioned individual covariates if available. Now, the bank's rewards are

$$\mathbb{E}_\delta[\delta I - (1-\delta)P] - \lambda\sqrt{\mathrm{Var}_\delta[\delta I - (1-\delta)P]}, \quad (2)$$

where P is the principal (without loss of generality, we let $P = 1$), I is interest, δ indicates whether the loan is repaid, and $\lambda \in \mathbb{R}_+$. The first term is expected profit and the second term is to risk aversion. We assume the goal of the bank is to maximize (2).

The individual rewards are $\mathbb{I}[a = 1]$, where $\mathbb{I}$ is the indicator function—i.e., the reward is 1 if the loan is offered and 0 if it is denied. Then, demographic parity says that loans should be given to majority and minority members with equal frequency (within an ϵ tolerance), and equal opportunity says that loans should

[2]Technically, our MDP is the belief MDP of the POMDP where the state p is unobserved.

be given to qualified majority and minority members at equal rates (we assume an applicant is qualified if their true probability of repaying satisfies $p \geq p_0$ for some $p_0 \in [0,1]$).

Separable MDPs. We focus primarily on MDPs where the fairness attribute is constant.

Definition 2.4. An MDP with states $S = Z \times \tilde{S}$ is *separable* if the transitions satisfy $P_{(z,\tilde{s}),a,(z',\tilde{s}')} = \delta_{z,z'}\tilde{P}_{\tilde{s},a,\tilde{s}'}$, where $\delta_{z,z'} = \mathbb{I}[z = z']$ and $\tilde{P} \in \mathbb{R}^{|\tilde{S}|\times|A|\times|\tilde{S}|}$ is a transition matrix.

That is, the transitions do not affect z, so the sensitive attribute $z \in Z$ does not change over time. This property is satisfied by many sensitive attributes (e.g., race and gender). Fairnes properties may not make sense when the sensitive attribute can change.

Existence and determinism. Unconstrained MDPs always have a deterministic optimal policy (Sutton and Barto, 2018); however, with a fairness constraint, this result may not hold:

Theorem 2.5. *There exists $\epsilon > 0$ and an MDP M such that $\Pi_{DP,\epsilon} = \varnothing$. There exists $\epsilon > 0$ and an MDP M such that π^* in (1) is not deterministic.*

We give a proof in Appendix B. For the following special case, we can prove existence of fair policies:

Definition 2.6. We say ρ is *state-independent* if for some $\tilde{\rho} \in \mathbb{R}^{|A|}$, we have $\rho_{s,a} = \tilde{\rho}_a$ for all $s \in S$.

Intuitively, this property captures settings where the decision-maker uses the state to choose actions (e.g., ability to repay), but the outcomes for the individuals only depend on whether the preferred action is taken (e.g., a loan offer). Our example M_{loan} has state-independent individual rewards.

Theorem 2.7. *If the individual rewards are state-independent, then (1) has a solution.*

Proof. Any policy π such that $\pi_{s,a} = \tilde{\pi}_a$ for all $s \in S$ and some $\tilde{\pi} \in \mathbb{R}^{|A|}$, satisfies $\pi \in \Pi_{\text{DP}}$. $\square$

Comparison to supervised learning. Our fairness definitions are natural generalizations of their counterparts for supervised learning. For example, in the supervised learning setting, demographic parity says that majority and minority members should, on average, be given positive outcomes at equal rates. Our extension to MDPs says that this property should hold on average across time—more precisely, averaged over $t \sim \text{Geometric}(\gamma)$, where γ is the discount factor.

Conversely, our constraint reduces to the supervised learning constraint setting when the state distribution is constant over time—i.e., $D^{(\pi,t)} = D$ is independent of t and π. To see this claim, note that a constant state distribution implies that $D^{(\pi)} = D$, so the state-action distribution is simply $\Lambda^{(\pi)}_{s,a} = D_s\pi_{s,a}$, and our MDP demographic parity constraint reduces to

$$\left|\mathbb{E}_{s\sim D_{\text{maj}},a\sim\pi_s}[\rho_{s,a}] - \mathbb{E}_{s\sim D_{\text{min}},a\sim\pi_s}[\rho_{s,a}]\right| \leq \epsilon.$$

In other words, the policy π should equalize the expected individual rewards for the majority and minority subpopulations on the initial (constant) state distribution. Finally, assuming the individual rewards are $\rho_{s,a} = 1$ for a positive outcome and $\rho_{s,a} = 0$ otherwise, then our constraint is equivalently

$$\left|\mathbb{P}_{s\sim D_{\text{maj}},a\sim\pi_s}[\hat{y} = 1] - \mathbb{P}_{s\sim D_{\text{min}},a\sim\pi_s}[\hat{y} = 1]\right| \leq \epsilon,$$

where $\hat{y} = \rho_{s,a}$ is the outcome, which is demographic parity for supervised learning (Hardt et al., 2016).

Additionally, we introduce individual rewards ρ, which may differ from the decision maker rewards R. This distinction also appears in the supervised learning setting if the loss function for the decision maker (used in the learning objective) differs from the loss function of the individual (used in the fairness constraint). For example, R may differ from ρ if the decision maker is risk-averse; then, the decision maker may offer too few loans to minorities if there is less historical information available for minorities. We believe this distinction is particularly important to explicitly model in the MDP setting, since dynamical effects can magnify the negative consequences of unfair decision making.

Importance of dynamics. Dynamics are important when current decisions do not immediately cause unfairness, but can affect the state distribution in a way that leads to unfair outcomes in the future. In our loan applicant example, there are two effects of decisions on the state distribution. First, there is a direct effect—e.g., denying loans can cause adverse outcomes on an applicant's financial situation. In M_{loan}, this effect is captured by the update $\beta_{t+1} = \beta_t + \tau$ when $a = 0$—i.e., the applicant's probability of repaying future loans decreases when they are denied a loan.

The second effect is indirect, and is related to the selective labels problem in sequential decision making (Lakkaraju et al., 2017; Bechavod et al., 2019). In particular, the bank only observes outcomes if they offer the applicant a loan. A key concern is that less historical information is available for minorities, leading to higher variance estimates of their ability to repay a loan. Thus, a risk-averse decision maker might conservatively deny loans to minorities, even if their expected rate of repaying loans is equal to that of majority members. The equal opportunity constraint forces the decision maker to give exploratory loans to avoid unfairly denying loans to an applicant for whom little data is available.

Algorithm 1 Algorithm for finite state, separable MDPs.

procedure LEARNFAIRPOLICY(Separable MDP M)

Compute the solution λ^* to the linear program

$$\begin{aligned}
&\operatorname*{arg\,max}_{\lambda\in\mathbb{R}^{|S|\times|A|}} \quad (1-\gamma)^{-1}\sum_{s\in S}\sum_{a\in A}\lambda_{s,a}R_{s,a} \\
&\text{subj. to} \quad \sum_{a\in A}\lambda_{s',a} = (1-\gamma)D_{s'} + \gamma\sum_{s\in S}\sum_{a\in A}\lambda_{s,a}P_{s,a,s'} \qquad (\forall s'\in S) \\
&\qquad\quad \left| p_{\text{maj}}^{-1}\sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A}\lambda_{(\text{maj},\tilde{s}),a}\rho_{(\text{maj},\tilde{s}),a} - p_{\text{min}}^{-1}\sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A}\lambda_{(\text{min},\tilde{s}),a}\rho_{(\text{min},\tilde{s}),a} \right| \le \epsilon
\end{aligned}$$

return π^*, where $\pi^*_{s,a} = \dfrac{\lambda^*_{s,a}}{\sum_{a'\in A}\lambda^*_{s,a'}}$

end procedure

3 Algorithm for Finite-State MDPs

We describe an algorithm for solving (1), which has strong theoretical guarantees (i.e., it solves (1) exactly in polynomial time). On the other hand, it makes strong assumptions—i.e., that M has finite state and action spaces. In Section 4, we describe a model-free algorithm that applies very generally (e.g., to continuous state and action spaces, or even non-separable MDPs), but lacks performance guarantees.

Our approach is based on the dual of the standard LP formulation of value iteration (Altman, 1999; Sutton and Barto, 2018). In particular, the objective and first set of constraints of the LP in Algorithm 1 form the dual. The last set of constraints in the LP in Algorithm 1 encodes demographic parity. These constraints exploit the separable structure of the underlying MDP. In particular, the component z of an initial state $s = (z, \tilde{s})$ does not change over time, so the value of z for s equals the value of z for the initial state $s_0 \sim D$. Thus, randomly sampling a state $s \sim D_z^{(\pi)}$ is equivalent to randomly sampling

$$s \sim D^{(\pi)} \mid \exists \tilde{s} \in \tilde{S} \ . \ s = (z, \tilde{s}).$$

Expanding the conditional probability, the probability of sampling $s \sim D_z^{(\pi)}$ is

$$\frac{D_s^{(\pi)}\mathbb{I}[\exists \tilde{s}\in\tilde{S} \ . \ s=(z,\tilde{s})]}{p_z}, \quad \text{where } p_z = \sum_{\tilde{s}\in\tilde{S}} D_{(z,\tilde{s})}.$$

It follows that

$$\rho_z^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}] = p_z^{-1}\sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A}\lambda_{(z,\tilde{s}),a}\rho_{s,a}. \quad (3)$$

The last set of constraints in the LP in Algorithm 1 uses (3) to encode demographic parity.

Theorem 3.1. *Algorithm 1 returns a solution π^* to (1) if and only if (1) is satisfiable.*

We give a proof in Appendix C. Note that Algorithm 1 runs in polynomial time.

Remark 3.2. We briefly compare our approach to algorithms for solving constrained MDPs. Existing approaches are also based on the dual of the LP for solving MDPs (Altman, 1999). Indeed, in the LP we use in Algorithm 1, the objective and the first constraint are taken from the dual. The second constraint, which encodes the fairness constraint, is novel—our key insight is that for separable MDPs, the fairness constraint can be expressed as a linear inequality over λ.

4 Algorithm for General MDPs

Next, we propose a general algorithm for solving (1). However, in general, the planning problem may be non-convex, so unlike Algorithm 1, this algorithm may converge to a local optimum.

Our algorithm relies on the cross-entropy (CE) method (Mannor et al., 2003; Hu et al., 2012), a heuristic for solving optimization problems. Suppose our policies $\pi_\theta \in \Pi$ are parameterized by $\theta \in \Theta$, and let a family $\mathcal{F}$ of probability distributions over Θ parameterized by $V \subseteq \mathbb{R}^d$. We use θ and π_θ interchangeably, e.g., $R^{(\theta)} = R^{(\pi_\theta)}$. In the unconstrained setting, CE aims to solve the following optimization problem:

$$v^* = \operatorname*{arg\,max}_{v\in V}\mathbb{E}_v[R^{(\theta)}], \quad (4)$$

where $\mathbb{E}_v = \mathbb{E}_{\theta\sim f_v}$. In other words, it aims to compute a distribution f_{v^*} that places high probability mass on θ with high cumulative expected reward $R^{(\theta)}$. Then, it returns a sample $\theta \sim f_{v^*}$. To solve (4), CE starts with initial parameters $v_0 \in V$. Then, on each iteration, it

Algorithm 2 Algorithm for general MDPs.

1: **procedure** GeneralLearnFairPolicy(MDP M, Iters r, Parameter samples n, Top n', Rollout samples m, Smoothing α, Tolerance σ)
2: $\quad \hat{\eta} \leftarrow \vec{0}$
3: $\quad$ **for** $k \in [1, ..., r]$ **do**
4: $\quad\quad$ Sample $\theta^{(1)}, ..., \theta^{(n)} \sim f_{m^{-1}(\hat{\eta})}$
5: $\quad\quad$ **for** $i \in [1, ..., n]$ **do**
6: $\quad\quad\quad \hat{R}^{(\theta^{(i)})} \xleftarrow{\sim m,T} R^{(\theta^{(i)})}$
7: $\quad\quad\quad \hat{\epsilon}^{(\theta^{(i)})} \xleftarrow{\sim m,T} |\rho_{\text{maj}}^{(\theta^{(i)})} - \rho_{\text{min}}^{(\theta^{(i)})}|$
8: $\quad\quad$ **end for**
9: $\quad\quad$ Sort $\{\theta^{(i)}\}_{i=1}^{n}$ in increasing $\hat{\epsilon}^{(\theta^{(i)})}$
10: $\quad\quad i' \leftarrow$ Largest i such that $\hat{\epsilon}^{(\theta^{(i)})} \leq (1-\sigma)\epsilon$
11: $\quad\quad$ **if** $n' \leq i'$ **then**
12: $\quad\quad\quad$ Sort $\{\theta^{(i)}\}_{i=1}^{i'}$ in decreasing $\hat{R}^{(\theta^{(i)})}$
13: $\quad\quad$ **end if**
14: $\quad\quad \hat{\eta} \leftarrow \alpha \cdot \frac{\frac{1}{n}\sum_{i=1}^{n'} \hat{R}^{(\theta^{(i)})}\Gamma(\theta^{(i)})}{\frac{1}{n}\sum_{i=1}^{n'} \hat{R}^{(\theta^{(i)})}} + (1-\alpha)\cdot\hat{\eta}$
15: $\quad$ **end for**
16: $\quad$ **if** $\hat{\epsilon}^{(\hat{\theta})} \leq \tilde{\epsilon}$, where $\hat{\theta} \sim f_{m^{-1}(\hat{\eta})}$ **then**
17: $\quad\quad$ **return** $\pi_{\hat{\theta}}$
18: $\quad$ **else**
19: $\quad\quad$ **return** $\varnothing$
20: $\quad$ **end if**
21: **end procedure**

updates the current parameters v_k to move "closer" to v^*. More precisely, the update is

$$v_{k+1} = \arg\max_{v \in V} D_{\text{KL}}(g_{k+1} \,\|\, f_v) \tag{5}$$
$$g_{k+1}(\theta') = \alpha \frac{R^{(\theta')}\mathbb{I}[R^{(\theta')} \geq \gamma_k] f_{v_k}(\theta')}{\mathbb{E}_{v_k}[R^{(\theta)}\mathbb{I}[R^{(\theta)} \geq \gamma_k]]} + (1-\alpha) f_{v_k}(\theta')$$

where γ_k satisfies $\Pr_{v_k}[R^{(\theta)} \geq \gamma_i] = \mu$. Here, $\alpha, \mu \in (0,1)$ are hyperparameters. Intuitively, the first term of g_i upweights θ' with large values of $R^{(\theta')}$ compared to f_{v_k}, both by directly weighting the probability of θ' by $R^{(\theta')}$, and furthermore by placing zero probability mass on the bottom $1-\mu$ fraction of the θ'. The second term of g_k is a "smoothing" term that makes the update incremental.

To enable efficient optimization of (5), we assume that $\mathcal{F}$ is a (natural) exponential family.

Definition 4.1. A family $\mathcal{F}$ of distributions over $\Theta \subseteq \mathbb{R}^d$ is an *exponential family* if, for a continuous $\Gamma : \Theta \to \mathbb{R}^d$, $f_v(\theta) = e^{v^\top \Gamma(\theta)}/Z(\theta)$, where $Z(\theta) = \int e^{v^\top \Gamma(\theta)} d\theta$.

We use the standard choice that $\mathcal{F}$ is the space of Gaussians. If $\mathcal{F}$ is an exponential family, then

$$v_{k+1} = m^{-1}(\eta_{k+1}) \tag{6}$$
$$\eta_{k+1} = \alpha \frac{\mathbb{E}_{v_k}[R^{(\theta)}\mathbb{I}[R^{(\theta)} \geq \gamma_k]\Gamma(\theta)]}{\mathbb{E}_{v_k}[R^{(\theta)}\mathbb{I}[R^{(\theta)} \geq \gamma_k]]} + (1-\alpha)\eta_k$$

where $m(v) = \mathbb{E}_v[\Gamma(\theta)]$ is the moment map (Hu et al., 2012). The CE algorithm approximates (6) by sampling rollouts $\zeta = ((s_0, a_0), ..., (s_{T-1}, a_{T-1}))$ according to π_θ. Then, it computes the estimate $R^{(\theta)} \approx \hat{R}^{(\theta)} = \frac{1}{m}\sum_{i=1}^{m} \hat{R}(\zeta^{(i)})$, where $\zeta^{(1)}, ..., \zeta^{(m)}$ are m sampled rollouts and $\hat{R}(\zeta) = \sum_{t=0}^{T-1} \gamma^t R_{s_t, a_t}$.

To estimate η_{k+1}, it takes n samples $\theta^{(1)}, ..., \theta^{(n)} \sim f_v$, and computes $\hat{R}^{(\theta^{(i)})}$ for each i. Then, it ranks $\theta^{(i)}$ in decreasing order of $\hat{R}^{(\theta^{(i)})}$, and discards all but the top $n' = \lceil n\mu \rceil$. It estimates the numerator in η_{k+1} as

$$\mathbb{E}_{v_k}[R^{(\theta)}\mathbb{I}[R^{(\theta)} \geq \gamma_k]\Gamma(\theta)] \approx \frac{1}{n}\sum_{i=1}^{n'} \hat{R}^{(\theta^{(i)})}\Gamma(\theta^{(i)}).$$

The denominator in η_{k+1} is estimated similarly.

Algorithm 2 computes this estimate of the update (6) assuming the condition on Line 16 is satisfied (as we discuss below, the check is needed to enforce the constraint that $\pi \in \Pi_{\text{DP},\epsilon}$. Line 6 of Algorithm 2 computes the estimates $\hat{R}^{(\theta^{(i)})}$ for samples $\theta^{(i)} \sim f_{v_k}$ for $i \in [n]$, and Line 14 estimates η_{k+1}. On Line 6 & 7, the notation $\xleftarrow{\sim m,T}$ means to estimate a quantity using m sampled rollouts $\zeta^{(1)}, ..., \zeta^{(m)}$ each of length T.

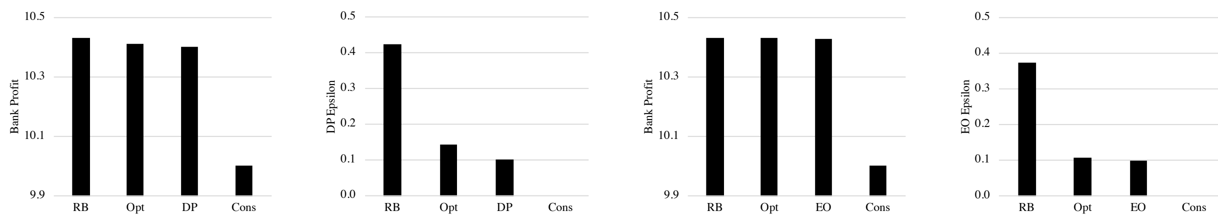
Finally, we adapt constrained cross-entropy (CCE), which extends CE to handle constraints (Wen and Topcu, 2018), to handle fairness constraints. Intuitively, CCE prioritizes policies where the constraint that $\pi \in \Pi_{\text{DP},\epsilon}$ is closer to holding, unless the constraint holds, in which case CCE prioritizes policies with higher cumulative expected reward. In particular, Algorithm 2 imposes this constraint by checking if $\hat{\theta}$ satisfies the constraint $\hat{\epsilon}^{(\hat{\theta})} \leq \epsilon$ in Line 16, where $\hat{\epsilon}^{(\hat{\theta})}$ is estimated from samples. Note that $\tilde{\epsilon}$ is used in place of ϵ to enforce the constraint even though $\hat{\epsilon}^{(\hat{\theta})}$ is inexact. The reason is that CCE relies on estimates $\hat{\epsilon}^{(\hat{\theta})}$ of $\epsilon^{(\hat{\theta})}$. These estimates are inexact since (i) they are estimated from samples, and (ii) they are estimated based on a finite time horizon (whereas $\epsilon^{(\hat{\theta})}$ is defined for an infinite horizon). To account for this error, we use $(1-\sigma)\epsilon$ (where $\sigma \in (0,1)$) in place of ϵ when checking the constraint on Line 16 of Algorithm 2.

We provide the following for Algorithm 2 (see Appendix D for a proof).

Theorem 4.2. *Assume that ρ_{max} is an upper bound on ρ (i.e., $\|\rho\|_\infty = \rho_{max}$ for all $z \in Z$). Let $\delta \in \mathbb{R}_+$ and $\sigma \in (0, 1/2]$ be given, and suppose that*

$$m \geq \frac{32\rho_{max}(1-\gamma)\log(4/\delta)}{\sigma\epsilon^2} \qquad T \geq \log\frac{4\rho_{max}}{\sigma^2\epsilon(1-\gamma)}.$$

Then, with probability at least $1-\delta$, we have $\pi_{\hat{\theta}} \in \Pi_{DP,\epsilon}$, where $\pi_{\hat{\theta}}$ is returned by Algorithm 2.



(a) (b) (c) (d)

Figure 1: Demographic parity (a) objective value, (b) constraint value, and equal opportunity (c) objective value, (d) constraint value, for race-blind (RB), demographic parity (DP) or equal opportunity (EO), optimistic (Opt), and conservative (Cons).

5 Reinforcement Learning

We discuss extensions to the setting where the MDP is initially unknown, and the goal is to ensure fairness while learning these quantities. We propose an approach to fairness when the transitions P are unknown but the initial state distribution D is known; reducing to the case of unknown D is standard (i.e., add a deterministic initial state s_0 and transition to an initial state according to D). Our goal is to ensure that with high probability, fairness holds for all time including during learning. We consider the episodic case where the system is reset after a fixed number of steps T, and take $\gamma = 1$. That is, a finite sequence of interactions is performed repeatedly—e.g., each new loan applicant is a new episode. We assume there are a fixed total number of episodes N, and the goal is to perform well on average; the doubling trick can be used to generalize to unknown or unbounded N (see p. 99 of Lattimore and Szepesvári).

A key challenge is how to design a fair policy we can use when the dynamics are unknown. Thus, we focus on the setting of state-independent individual rewards ρ, where we can ensure such a policy exists. In particular, we take π_0 to choose actions uniformly randomly—i.e., $\pi_0(s, a) = 1/|A|$ for all $s \in S$ and $a \in A$. Then, we are guaranteed that π_0 is fair. Furthermore, we are guaranteed that π_0 explores all states (assuming without loss of generality that we prune unreachable states)—i.e., letting $D^{(\pi)} = \frac{1}{T}\sum_{t=0}^{T-1} D^{(\pi,t)}$ and $\Lambda_{s,a}^{(\pi)} = D_s^{(\pi)}\pi_{s,a}$, where $D^{(\pi,t)}$ is defined as before, then there exists $\lambda_0 \in \mathbb{R}_+$ such that

$$\Lambda_{s,a}^{(\pi_0)} \geq \lambda_0 > 0 \qquad (\forall s \in S,\ a \in A)$$

We use explore-then-commit (Lattimore and Szepesvári). First, we explore using the conservative policy π_0 for N_0 episodes. Then, we estimate P using the observed state-action-state tuples (s, a, s') (i.e., transition to s' upon taking action a in state s):

$$\hat{P}_{s,a,s'} = \frac{\#\text{ observed tuples } (s, a, s')}{\#\text{ observed tuples } (s, a, s'') \text{ for some } s'' \in S}.$$

Finally, for the remaining $N - N_0$, it uses the optimal policy $\hat{\pi}$ computed as if $\hat{P}$ is the true transition matrix.

We prove a bound on the *regret*

$$\mathcal{R}(N) = \mathbb{E}\left[\sum_{n=1}^{N} R^{(\pi^*)} - R^{(\pi_n)}\right],$$

where the expectation is taken over the randomness of the observed tuples (s, a, s'), π^* is the optimal policy for known P that satisfies $\pi^* \in \Pi_{\mathrm{DP},\epsilon/4}$, and

$$\pi_n = \begin{cases} \pi_0 & \text{if } n \leq N_0 \\ \hat{\pi} & \text{otherwise} \end{cases}$$

$$N_0 = \frac{128T^4 \cdot |S|^2 \cdot R_{\max}^2 \cdot \log(2|S|^2|A|/\delta)}{\lambda_0^2\tilde{\epsilon}^2}.$$

is the policy our algorithm uses on episode n. We show that $\hat{\pi}$ is fair, and that given $\delta \in \mathbb{R}_+$, $\pi_n \in \Pi_{\mathrm{DP},\epsilon}$ for every $n \in [N]$ with probability at least $1 - \delta$.

Theorem 5.1. *Let $\epsilon, \delta \in \mathbb{R}_+$ be given. Assume that R_{max} is an upper bound on R (i.e., $\|R\|_\infty = R_{max}$) and on ρ. Let $\hat{M} = (S, A, D, \hat{P}, R, T)$, and $\hat{\pi}$ be the optimal policy for $\hat{M}$ in $\Pi_{DP,\epsilon/2}$ (i.e., the set of policies satisfying demographic parity for $\hat{M}$). Let $M = (S, A, D, P, R, T)$, and π^* be optimal for M in $\Pi_{DP,\epsilon/4}$. Then, $\hat{\pi} \in \Pi_{DP,\epsilon}$, and $\mathcal{R}(N) = O((N^{2/3} + 1/\epsilon^2)\log(1/\delta))$ with probability at least $1 - \delta$.*

We give a proof in Appendix E. Note that there is a gap between the fairness constraint of π^* (which is in $\Pi_{\mathrm{DP},\epsilon/4}$) and that of $\hat{\pi}$ (which is only in $\Pi_{\mathrm{DP},\epsilon}$)—i.e., we can only guarantee performance compared to a policy that satisfies a stricter level of fairness.

6 Experiments

We run simulations using our loan example from Section 2. We estimated parameters based on FICO score data (Hardt et al., 2016). We consider Whites to be majorities, and Blacks, Hispanics, and Asians to be minorities. For the initial distribution p_0, we first fit parameters the parameters of the prior $\mathrm{Beta}(\alpha_z, \beta_z)$ based on the data. Then, we take a fixed number of steps T_z using action $a = 1$ (i.e., offer loan) to force exploration. We choose $T_{\mathrm{maj}} > T_{\mathrm{min}}$ to capture the idea that less data is available for minorities. We also estimate the probability p_Z of being a minority from the data. Similar to (Hardt et al., 2016), we choose I so the bank makes a profit on the average applicant. We manually choose λ, τ, T_{maj}, and T_{min} based on intuition; see Appendix G for the values we chose. We focus on evaluation of Algorithm 2, and give additional experimental results in Appendix G.

Baselines that ignore dynamics. To demonstrate the importance of accounting for dynamics, we compare to two baselines that ignore dynamics when constraining fairness. The first optimistically pretends that actions do not affect the state distribution—i.e., $D^{(\pi,t)}$ does not change over time. In this case, for all $t > 0$, we have $D^{(\pi,t)} = D$, so $D^{(\pi)} = D$ for any π. Thus, we can let

$$\pi^* = \underset{\pi\in\Pi, c\in\mathbb{R}}{\arg\max}\, R^{(\pi)} \qquad (7)$$
$$\text{subj. to } \mathbb{E}_{s\sim D_z}\left[\sum_{a\in A} \pi_{s,a}\rho_{s,a}\right] = c \qquad (\forall z \in Z),$$

where $D_z = D \mid \exists \tilde{s} \in \tilde{S} \,.\, s_0 = (z, \tilde{s})$. We can solve (7) using a straightforward modification of Algorithm 2. This captures the supervised learning setting. Compared to our algorithm, this algorithm may learn a policy that is unfair but achieves higher reward.

The second conservatively assumes $D^{(\pi,t)}$ can change arbitrarily on each step. This baseline learns a fair policy, but it may achieve much lower reward. In this case, we restrict to policies π that satisfy

$$\mathbb{E}_{s\sim D'_{\mathrm{maj}}}\left[\sum_{a\in A} \pi_{s,a}\rho_{s,a}\right] = \mathbb{E}_{s\sim D'_{\mathrm{min}}}\left[\sum_{a\in A} \pi_{s,a}\rho_{s,a}\right] \qquad (8)$$
$$(\forall D' \in \Delta^{|S|}),$$

where $D'_z = D' \mid \exists \tilde{s} \in \tilde{S} \,.\, s = (z, \tilde{s})$, and Δ^n is the standard n-simplex. Note that D'_z is conditioned on $s = (z, \tilde{s})$ (i.e., the current state has sensitive attribute z) instead of $s_0 = (z, \tilde{s})$ (i.e., the initial state has sensitive attribute z); if M is separable, these two conditions are equivalent. Finally, note that D'_z is undefined if the conditional has zero probability according to D'; we implicitly omit such D' from (8).

The difficulty with (8) is the universal quantification over $D' \in \Delta^{|S|}$. For state-independent individual rewards, the conservative assumption is in fact equivalent to optimizing over state-independent policies—i.e., those of the form $\pi_{s,a} = \tilde{\pi}_a$, where $\tilde{\pi} \in \mathbb{R}^{|A|}$. Thus, we can apply a modified version of Algorithm 2 where we only learn state-independent policies.

Results for Algorithm 2. We ran Algorithm 2 to learn fair policies for both the demographic parity and equal opportunity constraints, using $\epsilon = 0.1$. For each constraint, we also use our optimistic and conservative baselines. We also consider a race-blind algorithm that is unconstrained but where π ignores the sensitive attribute $z \in Z$. The optimal policy is race-blind—the state is a sufficient statistic, so it captures all information needed to determine whether to offer a loan.

For demographic parity, Figure 1 (a) shows the reward achieved for the bank, and (b) shows the value of the fairness constraint—i.e., the smallest value of ϵ for which $\pi \in \Pi_{\mathrm{DP},\epsilon}$. As expected, race-blind achieves the highest reward (10.43), followed by the optimistic algorithm (10.41), and then Algorithm 2 (10.40). Finally, the conservative algorithm performs substantially worse than the others (10.00). However, race-blind achieves a very poor constraint value (0.42), as does the optimistic algorithm (0.14), which performs performs 43% worse than Algorithm 2 (0.10). The conservative algorithm achieves constraint value 0. For equal opportunity, Figure 1 (c) shows the bank reward, and (d) shows the value of the constraint. The bank's rewards are essentially the same for the race-blind algorithm, optimistic algorithm, and Algorithm 2 (10.43), but is substantially worse for the conservative algorithm (10.00). As with demographic parity, the constraint value for race-blind (0.37) is substantially worse than the others, but in this case optimistic (0.11) is fairly close to Algorithm 2 (0.10). The conservative algorithm achieves constraint value 0.

Discussion. Our results show that imposing demographic parity slightly reduces the bank's reward, but substantially increases fairness compared to the race-blind and optimistic algorithms. The latter models supervised learning—thus, our results show the importance of accounting for dynamics when ensuring fairness. We find similar (but weaker) trends for equal opportunity. Like prior work (Hardt et al., 2016), we find that demographic parity reduces the bank's rewards more than equal opportunity.

Unlike the static case (Hardt et al., 2016), our model has dynamic parameters. Time series data would be needed to estimate them; instead, we choose them manually. Also, (Hardt et al., 2016) uses the empirical CDF of the distribution over repayment probabilities p_0, whereas we assumed p_0 is a Beta distribution. Our goal is to understand the consequences of ignoring dynamics, not to study a real-world scenario.

7 Conclusion

We have proposed algorithms to learn fair policies that account for the dynamical effects, and have demonstrated the importance of accounting for these effects. There is much room for future work. One important direction is extending our results for the case where the initial MDP is unknown beyond explore-then-commit to obtain better regret guarantees. Another direction is theoretically analyzing the cost of fairness—e.g., what is the cost to the bank for imposing a fairness constraint, and how they can mitigate this cost by improving predictive power. Finally, reinforcement learning problems in practice are often offline—i.e., the goal is to learn from historical data and the algorithm does not have the opportunity to explore. Studying fairness in this context is an important problem.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This research was partially supported by NSF awards 1910769 and 1652113.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *ICML*, 2017.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Pranjal Awasthi, Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Beyond individual and group fairness. *arXiv preprint arXiv:2008.09490*, 2020.

Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Equal opportunity in online classification with partial feedback. *arXiv preprint arXiv:1902.02242*, 2019.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. *arXiv preprint arXiv:1909.09141*, 2019.

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*, pages 525–534, 2020.

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*, pages 214–226. ACM, 2012.

Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. 2019.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

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

Jiaqiao Hu, Ping Hu, and Hyeong Soo Chang. A stochastic approximation framework for a class of randomized optimization algorithms. *IEEE Transactions on Automatic Control*, 57(1):165–178, 2012.

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-Volume 70*, pages 1617–1626. JMLR. org, 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*, pages 325–333, 2016.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. AISTATS, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284. ACM, 2017.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *ICML*, 2018.

Shie Mannor, Reuven Y Rubinstein, and Yohai Gat. The cross entropy method for fast policy search. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 512–519, 2003.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Min Wen and Ufuk Topcu. Constrained cross-entropy method for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7461–7471, 2018.