arXiv:1902.02979v2 [cs.LG] 27 May 2019

Fair Decisions Despite Imperfect Predictions

Niki Kilbertus[†‡] Manuel Gomez-Rodriguez[♯]
Bernhard Schölkopf[†] Krikamol Muandet[†] Isabel Valera[†]

[†]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[‡]Department of Engineering, University of Cambridge, United Kingdom
[♯]Max Planck Institute for Software Systems, Kaiserslautern, Germany

Abstract

Consequential decisions are increasingly informed by sophisticated data-driven predictive models. For accurate predictive models, deterministic threshold rules have been shown to be optimal in terms of utility, even under a variety of fairness constraints. However, consistently learning accurate predictive models requires access to ground truth labels. Unfortunately, in practice, labels only exist conditional on certain decisions, which may have been made using a potentially imperfect decision policy. As a result, learned deterministic threshold rules are often suboptimal. Can we do better if we learn to decide rather than to predict? We first show that, if decisions are taken by a faulty deterministic policy, the observed labels are insufficient to improve it. Then, we describe how to avoid this undesirable behavior by directly learning stochastic decision policies that maximize utility under fairness constraints. Experiments on synthetic and real-world data illustrate the favorable properties of learning to decide in terms of utility and fairness.

1 Introduction

The use of machine learning models to assist consequential decision making—decision making which has significant consequences—is becoming common in a variety of critical applications. For example, in pretrial release decisions, a judge may consult a machine learning model estimating the probability that the defendant would reoffend upon release, to decide whether she grants bail or not. In loan decisions, a bank may decide whether or not to offer a loan on the basis of a machine learning model's estimate of the probability that the individual would repay the loan. In fraud detection, an insurance company may flag suspicious claims based on a machine learning model's estimate of the probability that the claim is fraudulent. In all these scenarios, the primary goal of the decision maker (bank, law court, or insurance company) is to take decisions that maximize a given utility function. In contrast, the goal of the machine learning model is to provide an accurate prediction of the outcome, referred to as ground truth label (in short, label).

In this context, there has been a flurry of work on computational mechanisms to ensure that machine learning models do not disproportionately harm particular demographic groups sharing one or more sensitive attributes, e.g., race or gender (Dwork et al., 2012; Feldman et al., 2015). However, most previous work does not distinguish between *decisions* and *label predictions* and, consequently, suggested an inherent trade-off between utility (or fairness) and prediction accuracy (Chouldechova, 2017; Kleinberg et al., 2017b). Only recently has the distinction been made explicit (Corbett-Davies et al., 2017; Kleinberg et al., 2017a; Mitchell et al., 2018; Valera et al., 2018). This recent line of work has also shown that, if a predictive model achieves perfect prediction accuracy, *deterministic threshold rules*, a natural class of decision policies in which decisions are derived deterministically from the predictive model simply by thresholding, achieve maximum utility under various fairness constraints. This lends support to focusing on deterministic threshold policies and seemingly justifies using predictions and decisions interchangeably.

However, in many practical scenarios, the decision determines whether a label is realized or not—if bail (a loan) is denied, there is not even an option for the individual to reoffend (pay back the loan). This problem has been referred to by Lakkaraju et al. (2017) as *selective labels*. As a consequence, the label data used to train predictive models often depends on the decisions taken and this likely leads to suboptimal performance of the model. Even worse, deterministic threshold rules using even slightly imperfect predictive models can be far from optimal (Woodworth et al., 2017). This negative result raises the following question: *Can we do better if we learn directly to decide rather than to predict?*

In this paper, we first articulate how the "learning to predict" approach fails in a utility maximization setting (with fairness constraints) that accommodates a variety of real-world applications, including those mentioned previously. We show that label data gathered under deterministic policies (e.g., prediction based threshold rules) are not sufficient to improve the underlying predictive model, nor the decision policy. Then, we demonstrate how to overcome this undesirable behavior using a particular family of stochastic policies, *exploring policies*, and introduce a simple gradient-based algorithm to learn such policies. Our approach effectively leverages the observed labels to find an exploring policy that maximizes utility subject to fairness constraints. Experiments on synthetic and real-world data corroborate our theoretical results, showing that *learning to predict* is inherently limited and inferior to *learning to decide* for consequential decision making under imperfect predictions. All code is publicly available at `https://github.com/nikikilbertus/fair-decisions`.

Related work. Our work is most closely related to a recent line of work that analyzes the long-term effects of consequential decisions informed by data-driven predictive models on underrepresented groups (Hu & Chen, 2018; Liu et al., 2018; Mouzannar et al., 2019; Tabibian et al., 2019). However, their focus is mainly on the evolution of several measures of well-being under a perfect predictive model, neglecting the data collection phase (Dimitrakakis et al., 2019; Holstein et al., 2018). In contrast, we focus on analyzing how to improve a suboptimal policy when labels exist only for positive decisions. More broadly, our work relates to the growing literature on fairness in machine learning, which mostly attempts to match various statistics of the predictive models across protected subgroups.

We build on previous work on counterfactual inference and policy learning (Athey & Wager, 2017; Ensign et al., 2018; Gillen et al., 2018; Heidari & Krause, 2018; Joseph et al., 2016; Jung et al., 2018; Kallus, 2018; Kallus & Zhou, 2018; Lakkaraju & Rudin, 2017). These works usually assume that, given a decision, the label is always observed and, moreover, that the decision may influence the label distribution. In contrast, in our setting the label exists only if the decision is positive and the decision does not influence the label distribution. Two notable exceptions are Kallus & Zhou (2018) and Ensign et al. (2018), which also consider limited feedback. To highlight the differences, Kallus & Zhou (2018) focuses on the design of unbiased estimates for several fairness measures, rather than learning a policy. Ensign et al. (2018) consider a sequential decision making setting, assuming a deterministic mapping between features and labels. This allows them to reduce the problem to the apple tasting problem (Helmbold et al., 2000). Remarkably, in this deterministic setting, they also conclude that the optimal policy should be stochastic.

Finally, our work also overlaps with previous work on contextual bandits (Dudík et al., 2011; Langford et al., 2008; Agarwal et al., 2014) and reinforcement learning (Jabbari et al., 2016; Sutton & Barto, 1998). This literature has also shown that, to learn accurate predictive models, stochastic policies are often necessary to ensure adequate exploration (Silver et al., 2014). This is in contrast with the fairness literature, where deterministic policies have been predominantly argued for (Corbett-Davies et al., 2017; Valera et al., 2018; Meyer et al., 2019). However, their problem setting differs fundamentally from ours and they do not account for fairness constraints nor the selective labels problem. A recent notable exception is Joseph et al. (2016), which initiated the study of fairness in multi-armed bandits, however, their fairness notion is orthogonal to the most popular notions of fairness, which we consider in our work, and ignore the selective labels problem.

2 Decision policies and imperfect predictive models

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the feature domain, $\mathcal{S} = \{0, 1\}$ the range of sensitive attributes,[1] and $\mathcal{Y} = \{0, 1\}$ the set of ground truth labels. We assume the standard sigma algebras on these spaces. A *policy* is a mapping $\pi : \mathcal{X} \times \mathcal{S} \to \mathcal{P}(\{0, 1\})$ that maps an individual's feature vector and sensitive attribute to a probability distributions over binary *decisions*. For each individual, $\boldsymbol{x}, s$ and y are sampled from a ground truth distribution $P(\boldsymbol{x}, s, y) = P(y \,|\, \boldsymbol{x}, s)P(\boldsymbol{x}, s)$. Each decision d is sampled from a given policy $d \sim \pi(d \,|\, \boldsymbol{x}, s)$, where we often write $\pi(\boldsymbol{x}, s)$ for $\pi(d \,|\, \boldsymbol{x}, s)$. The decision controls whether the label $y \sim P(y \,|\, \boldsymbol{x}, s)$ comes into existence and it is observed. For example, in loan decisions, the feature vector $\boldsymbol{x}$ may include an individual's salary, education, or credit history; the sensitive attribute s may indicate sex (e.g., male or female); the label y indicates whether an individual repays a loan ($y = 1$) or defaults ($y = 0$) *upon receiving it*; and the decision d specifies whether the individual receives a loan ($d = 1$) or not ($d = 0$).

Inspired by Corbett-Davies et al. (2017), we measure the *(immediate) utility* as the expected overall profit provided by the policy with respect to the distribution P, i.e.,

$$u_P(\pi) := \mathbb{E}_{\boldsymbol{x},s,y\sim P,\, d\sim\pi(\boldsymbol{x},s)}\left[y\, d - c\, d\right] = \mathbb{E}_{\boldsymbol{x},s\sim P}\left[\pi(d = 1 \,|\, \boldsymbol{x}, s)(P(y = 1 \,|\, \boldsymbol{x}, s) - c)\right], \tag{1}$$

where $c \in (0, 1)$ may reflect economic considerations of the decision maker. For example, in a loan scenario, the utility gain is $(1 - c)$ if a loan is granted and repaid, $-c$ if a loan is granted but the individual defaults, and zero if the loan is not granted. One could think of adding a term for negative decisions of the form $g(y)(1 - d)$ for some given definition of g, however, we would not be able to compute such a term due to the selective labels, except for constant g. Therefore, without loss of generality, we assume that $g(y) = 0$ for all y, because any constant can easily be absorbed in our framework. For fairness considerations, we define the *(immediate) f-benefit for group $s \in \{0, 1\}$* with respect to the distribution P by

$$b^s_P(\pi) := \mathbb{E}_{\boldsymbol{x},y\sim P(\boldsymbol{x},y \,|\, s),\, d\sim\pi(\boldsymbol{x},s)}[f(d, y)], \quad \text{with } f : \{0, 1\} \times \{0, 1\} \to \mathbb{R}.$$

[1]For simplicity, we assume the sensitive attribute to be binary, potentially resulting in inadequate binary gender or race assignments. However, our work can be easily generalized to categorical sensitive attributes $s \in \{0, \ldots, k - 1\}$ with $k > 2$.

Note that various common fairness criteria can be expressed as $b^0_P(\pi) = b^1_P(\pi)$ for different choices of f, e.g., *demographic parity* (or no disparate impact) (Feldman et al., 2015) amounts to $f(d, y) = d$ and *equality of opportunity* (Hardt et al., 2016) amounts to $f(d, y) = d \cdot y$.

Under perfect knowledge of the conditional $P(y \,|\, \boldsymbol{x}, s)$, the policy maximizing the above utility subject to the group benefit fairness constraint $b^0_P(\pi) = b^1_P(\pi)$ is a deterministic threshold rule (Corbett-Davies et al., 2017):[2]

$$\pi^*(d = 1 \,|\, \boldsymbol{x}, s) = \mathbf{1}[P(y = 1 \,|\, \boldsymbol{x}, s) \geq c_s], \tag{2}$$

where now we have to allow for group specific cost factors c_0, c_1 such that $b^0_P(\pi) = b^1_P(\pi)$. Without fairness constraints, we simply have $c_1 = c_2 = c$. However, as discussed in Woodworth et al. (2017), in practice, we typically do not have access to the true conditional distribution $P(y \,|\, \boldsymbol{x}, s)$, but instead to an imperfect predictive model $Q(y \,|\, \boldsymbol{x}, s)$ trained on a finite training set. Such a predictive model can similarly be used to implement a deterministic threshold rule as:

$$\pi_Q(d = 1 \,|\, \boldsymbol{x}, s) = \mathbf{1}[Q(y = 1 \,|\, \boldsymbol{x}, s) \geq c] \tag{3}$$

Here, the predictor $Q(y = 1 \,|\, \boldsymbol{x}, s) \approx P(y = 1 \,|\, \boldsymbol{x}, s) - \delta_s$, with $\delta_s = c_s - c$, directly incorporates the fairness constraint, i.e., it is trained to maximize predictive power subject to the fairness constraint. In this context, Woodworth et al. (2017) have shown that this approach often leads to better performance than post-processing a potentially unfair predictor as proposed by Hardt et al. (2016). Unfortunately, they have also shown that, because of the mismatch between $Q(y = 1 \,|\, \boldsymbol{x}, s)$ and $P(y = 1 \,|\, \boldsymbol{x}, s) - \delta_s$, the resulting policy π_Q will usually still be suboptimal in terms of both utility and fairness. To make things worse, due to the selective labeling, the data points $\boldsymbol{x}, s, y$ observed under a given policy π_0 are not i.i.d. samples from the ground truth distribution $P(\boldsymbol{x}, s, y)$, but instead from the weighted distribution

$$P_{\pi_0}(\boldsymbol{x}, s, y) \propto P(y \,|\, \boldsymbol{x}, s)\, \pi_0(d = 1 \,|\, \boldsymbol{x}, s)\, P(\boldsymbol{x}, s). \tag{4}$$

Consequently, if π_0 is not perfect, i.e., $\pi_0 \neq \pi^*$, the standard learning theory results for the empirical risk do not hold (even asymptotically), which may also be one reason for a common observation in fairness, namely that predictive errors are often systematically larger for minority groups (Angwin et al., 2016). In the remainder, we will say that the distributions $P_{\pi_0}(\boldsymbol{x}, s, y)$ and $P_{\pi_0}(\boldsymbol{x}, s)$ are *induced* by the policy π_0. In the next section, we study how to learn the optimal policy, potentially subject to fairness constraints, if the data is collected from an initial faulty policy π_0.

3 Deterministic vs stochastic policies

Consider a class of policies Π, within which we want to maximize utility, as defined in eq. (1) subject to the group benefit fairness constraint $b^0_P(\pi) = b^1_P(\pi)$. Then, we formulate this as an unconstrained optimization problem with an additional penalty term

$$\text{maximize} \quad v_P(\pi) := u_P(\pi) + \frac{\lambda}{2}(b^0_P(\pi) - b^1_P(\pi))^2 \quad \text{over } \pi \in \Pi \tag{5}$$

under the assumption that we do not have access to samples from the ground truth distribution $P(\boldsymbol{x}, s, y)$, which $u_P(\pi)$ and $b^s_P(\pi)$ depend on. Instead we only have access to samples from a

[2] Here, $\mathbf{1}[\bullet]$ is 1 if the predicate $\bullet$ is true and 0 otherwise.

distribution $P_{\pi_0}(\boldsymbol{x}, s, y)$ induced by a given initial policy π_0 as in eq. (4). We first analyze this problem for deterministic threshold rules, before considering general deterministic policies, and finally move on to general stochastic policies.

3.1 Deterministic policies

Assume the initial policy π_0 is a given deterministic threshold rule and Π is the set of all deterministic threshold rules, which means that each $\pi \in \Pi$ (and π_0) is of the form eq. (3) for some predictive model $Q(y \,|\, \boldsymbol{x}, s)$. Given a hypothesis class of predictive models $\mathcal{Q}$, we reformulate eq. (5) to

$$\text{maximize} \quad v_P(\pi_Q) := u_P(\pi_Q) + \frac{\lambda}{2}(b_P^0(\pi_Q) - b_P^1(\pi_Q))^2 \quad \text{over } Q \in \mathcal{Q} \tag{6}$$

where the utility and the benefits for $s \in \{0, 1\}$ are simply $u_P(\pi_Q) = \mathbb{E}_{\boldsymbol{x},s,y\sim P}[\mathbf{1}[Q(y = 1 \,|\, \boldsymbol{x}, s) \geq c](y - c)]$ and $b_P^s(\pi_Q) = \mathbb{E}_{\boldsymbol{x},s,y\sim P}[f(\mathbf{1}[Q(y = 1 \,|\, \boldsymbol{x}, s) \geq c], y)]$. Note that eq. (5) has a unique optimum π^*. Therefore, if $\pi^* \in \Pi$ (the set of all deterministic threshold rules), eq. (6) will also reach this optimum if $\mathcal{Q}$ is rich enough, however, the optimal predictor Q^* may not be unique. Since the utility and the benefits are not sensitive to the precise values of $Q(y \,|\, \boldsymbol{x}, s)$ above or below c, we may have that a $Q \in \mathcal{Q}$ with $Q \neq Q^*$ may approximate $P(y = 1 \,|\, \boldsymbol{x}, s) - \delta_s$ more accurately than Q^*.

Therefore, if we only have access to samples from the distribution P_{π_0} induced by some $\pi_0 \neq \pi^*$, we may choose to simply learn a predictive model $Q_0^* \in \mathcal{Q}$ that maximizes the objective $v_{P_{\pi_0}}(\pi_Q)$, where the utility and the benefits are computed with respect to the induced distribution P_{π_0}. However, the following negative result shows that, under mild conditions, Q_0^* would lead to a suboptimal deterministic threshold rule:

Proposition 1. *If there exists a subset $\mathcal{V} \subset \mathcal{X} \times \mathcal{S}$ of positive measure under P such that $P(y = 1 \,|\, \mathcal{V}) \geq c$ and $P_{\pi_0}(y = 1 \,|\, \mathcal{V}) < c$, then there exists a maximum $Q_0^* \in \mathcal{Q}$ of $v_{P_{\pi_0}}$ such that $v_P(\pi_{Q_0^*}) < v_P(\pi_{Q^*})$.*

Proof. First, note that any deterministic policy π is fully characterized by the sets $W_d(\pi) = \{(\boldsymbol{x}, s) \,|\, \pi(d = 1 \,|\, \boldsymbol{x}, s) = d\}$ for $d \in \{0, 1\}$. For a deterministic threshold rule π_Q, we write $W_d(Q) = \{(\boldsymbol{x}, s) \,|\, \mathbf{1}[Q(y = 1 \,|\, \boldsymbol{x}, s) > c] = d\} = W_d(\pi_Q)$. By definition, we have that $v(\pi_Q) \leq v(\pi_{Q^*})$. We note that whenever the symmetric difference between the sets $W_d(Q)$ and $W_d(Q^*)$, $W_d(Q)\Delta W_d(Q^*)$, has positive inner measure (induced by P) for $d \in \{0, 1\}$ and a $Q \in \mathcal{Q}$, we have $v(\pi_Q) \neq v(\pi_{Q^*})$ and thus $v(\pi_Q) < v(\pi_{Q^*})$. Thus it only remains to show that $W_d(Q^*)\Delta W_d(Q_0^*)$ has positive inner measure for $d \in \{0, 1\}$. Since $P(y = 1 \,|\, \mathcal{V}) \geq c$ by assumption, we have $\mathcal{V} \subset W_1(Q^*)$. At the same time, because of $P_{\pi_0}(y = 1 \,|\, \mathcal{V}) < c$ by assumption, we have $\mathcal{V} \cap W_1(\pi_0) = \emptyset$. Finally, we note that for any $Q \in \mathcal{Q}$, we have that $v_{P_{\pi_0}}(Q) = v_{P_{\pi_0}}(Q \cdot \chi_{W_1(\pi_0)})$, where $\chi_\bullet$ is the indicator function on the set $\bullet$. Therefore, we can choose a maximum Q_0^* maximizing $v_{P_{\pi_0}}$ such that $W_1(Q_0^*) \subset W_1(\pi_0)$ and thus $\mathcal{V} \cap W_1(Q_0^*) = \emptyset$. Therefore $\mathcal{V} \subset W_1(Q_0^*)\Delta W_1(Q^*)$ and $\mathcal{V}$ has positive measure under P by assumption. Thus $W_d(Q_0^*)\Delta W_d(Q^*)$ has positive inner measure and we conclude $v_P(\pi_{Q_0^*}) < v_P(\pi_{Q^*})$. □

Supplementing this result, we will now prove that—in certain situations—a sequence of deterministic threshold rules, where each threshold rule is of the form of eq. (3) and its associated predictive model is trained using the data gathered through the deployment of previous threshold rules, fails to recover the optimal policy despite it being in the hypothesis class. To this end, we consider a *sequential* *policy learning task*, which is given by a tuple $(\pi_0, \Pi', \mathcal{A})$, where: a) $\Pi' \subset \Pi$ is the hypothesis class of policies, b) $\pi_0 \in \Pi'$ is the initial policy, and c) $\mathcal{A} : \Pi' \times (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^+ \to \Pi'$ is an update rule.[3] The update rule $\mathcal{A}$ takes an existing policy π_t within the class of allowed policies and a dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^n$ of $n \in \mathbb{N}_{>0}$ examples and produces an updated policy π_{t+1}, which typically aims to improve the policy in terms of the objective function $v_P(\pi)$ in eq. (5). In our setting, the dataset $\mathcal{D}$ is collected through the deployment of previous policies, i.e., from a mixture of the distributions $P_{\pi_\tau}(\boldsymbol{x}, s, y)$ with $\tau \leq t$.

To introduce useful notation and terminology, note that any deterministic threshold policy π is fully characterized by the sets $W_d(\pi) := \{(\boldsymbol{x}, s) \,|\, \pi(\boldsymbol{x}, s) = d\}$ for $d \in \{0, 1\}$, i.e., we can partition the space $\mathcal{X} \times \mathcal{S} = W_0(\pi) \cup W_1(\pi)$ into negative and positive decisions. Then, we say an update rule is *non-exploring on* $\mathcal{D}$ iff $W_0(\mathcal{A}(\pi, \mathcal{D})) \subset W_0(\pi)$. Intuitively, this means that no individual who has received a negative decision under the old policy π would have received a positive decision under the new policy produced by $\mathcal{A}(\pi, \mathcal{D})$). Remarkably, common learning algorithms for classification, such as gradient boosted trees are *error based*, i.e., they only change the decision function if they make errors on the training set. As a result, they lead to non-exploring update rules on $\mathcal{D}$ whenever the error is zero, i.e., $\sum_{(\boldsymbol{x},s,y)\in\mathcal{D}} \mathbf{1}[\pi(\boldsymbol{x}, s) \neq y] = 0$.

Proposition 2. *Let $(\pi_0, \Pi', \mathcal{A})$ be a sequential policy learning task, where $\Pi' \subset \Pi$ are deterministic threshold policies based on a class of predictive models, and let the initial policy be more strict than the optimal one, i.e., $W_0(\pi_0) \supsetneq W_0(\pi^*)$. If $\mathcal{A}$ is non-exploring on any i.i.d. sample $\mathcal{D} \sim P_{\pi_t}(\boldsymbol{x}, s, y)$ with probability at least $1 - \delta_t$ for all $t \in \mathbb{N}$, then $Pr[\pi_T \neq \pi^*] > 1 - \sum_{t=0}^{T} \delta_t$ for any $T \in \mathbb{N}$.*

Proof. At each step, we have

$$\Pr[\pi_t = \pi^*] = \Pr[W_0(\pi_t) = W_0(\pi^*)] \leq \Pr[W_0(\pi_t) \supset W_0(\pi^*)] \leq \delta_t + \Pr[\pi_{t-1} = \pi^*].$$

By the assumption that $\pi_0 \neq \pi^*$, we recursively get $\Pr[\pi_t = \pi^*] \leq \sum_{i=0}^{t} \delta_i$ which concludes the proof. □

We can thus conclude that, for error based learning algorithms under no fairness constraints, learning within deterministic threshold policies is guaranteed to fail. Intuitively, even though the optimal policy lies within the set of deterministic threshold policies, it cannot easily be approximated within this set when starting from a suboptimal predictive model.

Corollary 3. *A deterministic threshold policy $\pi \neq \pi^*$ with $\Pr[\pi(\boldsymbol{x}, s) \neq y] = 0$ under P will never converge to π^* under an error based learning algorithm for the underlying predictive model.*

Proof. Since error based learning algorithms lead to non-exploring policies whenever $\sum_{(\boldsymbol{x},s,y)\in\mathcal{D}} \mathbf{1}[\pi(\boldsymbol{x}, s) \neq y] = 0$, using the assumption $\Pr[\pi(\boldsymbol{x}, s) \neq y] = 0$, we can use Proposition 2 with $\delta_t = 0$ for all $t \in \mathbb{N}$. □

While we have focused on deterministic threshold rules, our results readily generalize to *all* deterministic policies, because we can always express any arbitrary deterministic policy π using eq. (3), e.g., for $Q(y = 1 \,|\, \boldsymbol{x}, s) = \mathbf{1}[\pi(d = 1 \,|\, \boldsymbol{x}, s) = 1]$ we have trivially $\pi_Q = \pi$. To conclude, if we can only observe the outcomes of previous decisions taken by a deterministic initial policy π_0, these outcomes may be insufficient to find the (fair) deterministic policy that maximizes utility.

[3] For a set $\mathcal{V}$, we write $\mathcal{V}^+ := \bigcup_{i=1}^{\infty} \mathcal{V}^i$.

3.2 Stochastic policies

To overcome the undesirable behavior exhibited by deterministic policies discussed in the previous section, one could just use a fully randomized initial policy $\pi_0(d \mid \boldsymbol{x}, s)$, where $\pi_0(d = 1 \mid \boldsymbol{x}, s) = 1/2$ for all $(\boldsymbol{x}, s)$. It readily follows from eq. (4) that samples from the induced distribution P_{π_0} are i.i.d. samples from the ground truth distribution, $P_{\pi_0} = P$. As a result, if the hypothesis class of predictive models $\mathcal{Q}$ is rich enough, we could learn the optimal policy π^* from data gathered using the policy π_0. However, in practice, using a fully randomized initial policy is unacceptable in terms of utility—it would entail offering (releasing) loans (defendants) by a fair coin flip until sufficient data has been collected. Fortunately, we will show next that an initial stochastic policy does not need to be fully randomized to be able to learn the optimal policy. We only need to choose an initial policy π_0 such that $\pi_0(d = 1 \mid \boldsymbol{x}, s) > 0$ on any measurable subset of $\mathcal{X} \times \mathcal{S}$ with positive probability under P, a requirement that is more acceptable for the decision maker in terms of initial utility. We refer to any policy with this property as an *exploring* policy.[4] For any exploring policy π_0, we can compute the utility in eq. (1) and the group benefits for $s \in \{0, 1\}$ using inverse propensity score weighting:

$$u_{P_{\pi_0}}(\pi, \pi_0) := \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0}\\ d\sim\pi(\boldsymbol{x},y)}}\left[\frac{d(y-c)}{\pi_0(d=1 \mid \boldsymbol{x}, s)}\right] \text{ and } b^s_{P_{\pi_0}}(\pi, \pi_0) := \mathbb{E}_{\substack{\boldsymbol{x},s,y\sim P_{\pi_0}\\ d\sim\pi(\boldsymbol{x},y)}}\left[\frac{f(d,y)}{\pi_0(d=1 \mid \boldsymbol{x}, s)}\right]. \quad (7)$$

Crucially, even though $u_P(\pi) = u_{P_{\pi_0}}(\pi, \pi_0)$ and $b^s_P(\pi) = b^s_{P_{\pi_0}}(\pi, \pi_0)$, the expectations are with respect to the induced distribution $P_{\pi_0}(\boldsymbol{x}, s, y)$, yielding the following positive result.

Proposition 4. *Let Π be the set of exploring policies and let $\pi_0 \in \Pi \setminus \{\pi^*\}$. Then,*

$$v(\pi^*) = \sup_{\pi\in\Pi\setminus\{\pi^*\}} \Big\{ u_{P_{\pi_0}}(\pi, \pi_0) + \frac{\lambda}{2}\big(b^0_{P_{\pi_0}}(\pi, \pi_0) - b^1_{P_{\pi_0}}(\pi, \pi_0)\big)^2 \Big\}.$$

Proof. We already know that the supremum is upper bounded by $v(\pi^*)$, i.e., it suffices to construct a sequence of policies $\{\pi_n\}_{n\in\mathbb{N}_{>0}} \subset \Pi \setminus \{\pi^*\}$ such that $v(\pi_n) \to v(\pi^*)$ for $n \to \infty$. Using notation from the proof of Proposition 1, we define

$$\pi_n(d = 1 \mid \boldsymbol{x}, s) := \begin{cases} 1 & \text{if } (\boldsymbol{x}, s) \in W_1(\pi^*) \\ \frac{1}{n} & \text{otherwise.} \end{cases}$$

It is clear that π_n is exploring, i.e., $\pi_n \in \Pi$, for all $n \in \mathbb{N}_{>0}$ as well as that $\pi_n \neq \pi^*$. To compute

$$\lim_{n\to\infty} v_{P_{\pi_0}}(\pi_n, \pi_0) = \lim_{n\to\infty} \left(u_{P_{\pi_0}}(\pi_n, \pi_0) + \frac{\lambda}{2}\big(b^0_{P_{\pi_0}}(\pi_n, \pi_0) - b^1_{P_{\pi_0}}(\pi_n, \pi_0)\big)^2 \right)$$

[4]A policy π is exploring, iff the true distribution P is absolutely continuous w.r.t. the induced distribution P_π. In simple words, the distribution from which data are collected must not ignore regions where the true distribution puts mass.

we look at the individual limits. For the utility we have

$$\begin{aligned}
\lim_{n\to\infty} u_{P_{\pi_0}}(\pi_n,\pi_0) &= \lim_{n\to\infty} \mathbb{E}_{\boldsymbol{x},s,y\sim P_{\pi_0}(\boldsymbol{x},s,y)}\left[\frac{\pi_n(d=1\,|\,\boldsymbol{x},s)}{\pi_0(d=1\,|\,\boldsymbol{x},s)}(y-c)\right] \\
&= \int_{W_1(\pi^*)} \frac{P(y=1\,|\,\boldsymbol{x},s)-c}{\pi_0(d=1\,|\,\boldsymbol{x},s)}\,dP_{\pi_0}(\boldsymbol{x},s) + \\
&\quad \lim_{n\to\infty}\frac{1}{n}\underbrace{\int_{W_1(\pi^*)^\complement} \frac{P(y=1\,|\,\boldsymbol{x},s)-c}{\pi_0(d=1\,|\,\boldsymbol{x},s)}\,dP_{\pi_0}(\boldsymbol{x},s)}_{=:C_1 \text{ with } |C_1|<\infty \text{ for any given exploring } \pi_0\in\Pi} \\
&= \int_{W_1(\pi^*)} (y-c)\,dP(\boldsymbol{x},s,y) + \lim_{n\to\infty}\frac{C_1}{n} \\
&= u_P(\pi^*).
\end{aligned}$$

Similarly, for the benefit terms with $f(d,y)=d$ or $f(d,y)=d\cdot y$ we have for $s\in\{0,1\}$

$$\begin{aligned}
\lim_{n\to\infty} b^s_{P_{\pi_0}}(\pi_n,\pi_0) &= \mathbb{E}_{\boldsymbol{x},y\sim P_{\pi_0}(\boldsymbol{x},y\,|\,s)}\left[\frac{f(\pi_n(d=1\,|\,\boldsymbol{x},s),y)}{\pi_0(d=1\,|\,\boldsymbol{x},s)}\right] \\
&= \int_{W_1(\pi^*)} \frac{f(1,P(y=1\,|\,\boldsymbol{x},s))}{\pi_0(d=1\,|\,\boldsymbol{x},s)}\,dP_{\pi_0}(\boldsymbol{x}\,|\,s) + \\
&\quad \lim_{n\to\infty}\frac{1}{n}\underbrace{\int_{W_1(\pi^*)^\complement} \frac{f(1,P(y=1\,|\,\boldsymbol{x},s))}{\pi_0(d=1\,|\,\boldsymbol{x},s)}\,dP_{\pi_0}(\boldsymbol{x}\,|\,s)}_{=:C_2^s \text{ with } |C_2^s|<\infty \text{ for any given exploring } \pi_0\in\Pi} \\
&= \int_{W_1(\pi^*)} f(1,y)\,dP(\boldsymbol{x},y\,|\,s) + \lim_{n\to\infty}\frac{C_2^s}{n} \\
&= b^s_P(\pi^*).
\end{aligned}$$

Because all the limits are finite, via the rules for sums and products of limits we get

$$\begin{aligned}
\lim_{n\to\infty} v_{P_{\pi_0}}(\pi_n,\pi_0) &= \lim_{n\to\infty} u_{P_{\pi_0}}(\pi_n,\pi_0) + \frac{\lambda}{2}(\lim_{n\to\infty} b^0_{P_{\pi_0}}(\pi_n,\pi_0) - \lim_{n\to\infty} b^1_{P_{\pi_0}}(\pi_n,\pi_0))^2 \\
&= u_P(\pi^*) + \frac{\lambda}{2}(b^0_P(\pi^*) - b^1_P(\pi^*))^2 \\
&= v_P(\pi^*)
\end{aligned}$$

□

This shows that—unlike within deterministic threshold models—within exploring policies we can learn the optimal policy using only data from an induced distribution. Finally, we would like to highlight that not all exploring policies may be (equally) acceptable to society. For example, in lending scenarios without fairness constraints (i.e., $\lambda=0$), it may appear wasteful to still deny a loan with probability greater than zero to individuals who are believed to repay by the current model. In those cases, one may like to consider exploring policies that, given sufficient evidence, decide $d=1$ deterministically, i.e., $\pi_0(d=1\,|\,\boldsymbol{x},s)=1$ for some values of $\boldsymbol{x},s$.

4 Learning exploring policies

In this section, we exemplify Proposition 4 via a simple, yet practical, gradient-based algorithm to find the solution to eq. (5) within a (differentiable) parameterized class of exploring policies $\Pi(\Theta)$ using data gathered by a given, already deployed, exploring policy π_0. While our algorithm works for any differentiable class of exploring policies, here we consider two examples of exploring policy classes in particular. First, the *logistic policy*, which is given by

$$\pi_{\boldsymbol{\theta}}(d = 1 \,|\, \boldsymbol{x}, s) = \sigma(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}) \in (0, 1),$$

where $\sigma(a) := \frac{1}{1+\exp(-a)}$ is the logistic function, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ are the model parameters, and $\boldsymbol{\phi} : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}^m$ is a fixed feature map. Second, the *semi-logistic policy*, which deterministically approves examples believed to contribute positively to the utility by the current model and only explores stochastically on the remaining ones, i.e.,

$$\tilde{\pi}_{\boldsymbol{\theta}}(d = 1 \,|\, \boldsymbol{x}, s) = \mathbf{1}\left[\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} \geq 0\right] + \mathbf{1}\left[\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} < 0\right] \, \sigma\left(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}\right) \in (0, 1].$$

More specifically, we rely on stochastic gradient ascent (SGA) (Kiefer et al., 1952) to learn the parameters of the new policy, i.e., $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \nabla_{\boldsymbol{\theta}} v_P(\pi_{\boldsymbol{\theta}})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i}$, where $\nabla_{\boldsymbol{\theta}} v_P(\pi_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} u_P(\pi_{\boldsymbol{\theta}}) - \lambda(\nabla_{\boldsymbol{\theta}} b_0(\pi_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} b_1(\pi_{\boldsymbol{\theta}}))$, and $\alpha_i > 0$ is the learning rate at step $i \in \mathbb{N}$. With the reweighting from eq. (7) and the log-derivative trick (Williams, 1992), we can compute the gradient of the utility and the benefits as

$$\nabla_{\boldsymbol{\theta}} u_P(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\substack{\boldsymbol{x},s,y \sim P_{\pi_0} \\ d \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{x},s)}}\left[\frac{d\,(y - c) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}}{\pi_0(d = 1 \,|\, \boldsymbol{x}, s)}\right] \text{ and } \nabla_{\boldsymbol{\theta}} b_P^s(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\substack{\boldsymbol{x},s,y \sim P_{\pi_0} \\ d \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{x},s)}}\left[\frac{f(d, y) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}}{\pi_0(d = 1 \,|\, \boldsymbol{x}, s)}\right] \quad (8)$$

where $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} := \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(d \,|\, \boldsymbol{x}, s)$ is often referred as the score function (Hyvärinen, 2005). Thereby, our implementation resembles the REINFORCE algorithm with horizon one.

Unfortunately, the above procedure has two main drawbacks. First, it may require on an abundance of data drawn from P_{π_0}, which can be unacceptable in terms of utility since π_0 may be far from optimal and should not be deployed for too long. Second, if $\pi_0(d = 1 \,|\, \boldsymbol{x}, s)$ is small in a region where $\pi_{\boldsymbol{\theta}}$ often takes positive decisions, one may expect that an empirical estimate of the above gradient will have high variance, due to similar arguments as in weighted inverse propensity scoring Sutton & Barto (1998). On the other hand, in most practical applications, online learning algorithms that update their model after every single decision are rather impractical. Typically, a fixed model will be deployed for a certain period of time, before it is updated using the data collected within this period. This is also a natural mode of operation for predictive models in real-world applications.

To overcome these drawbacks, we build two types of sequences of policies $\{\pi_{\boldsymbol{\theta}_t}\}_{t=0}^T$: a) the *iterative sequence* $\pi_{t+1} := \mathcal{A}(\pi_t, \mathcal{D}^t)$ with $\mathcal{D}^t \sim P_{\pi_t}(\boldsymbol{x}, s, y)$, where only the data gathered by the immediately previous policy are used to update the current policy; and b) the *aggregated sequence* $\pi_{t+1} := \mathcal{A}(\pi_t, \bigcup_{i=0}^t \mathcal{D}^i)$ with $\mathcal{D}^i \sim P_{\pi_i}(\boldsymbol{x}, s, y)$, where the data gathered by all previous policies are used to update the current policy. Appendix A presents the overall training procedure. Finally, we would like to highlight that we opt for the simple SGA approach on (semi-)logistic policies over, e.g., contextual bandits algorithms, because it provides a direct and fairer comparison with prediction based decision policies (e.g., logistic regression), also commonly trained via SGA.

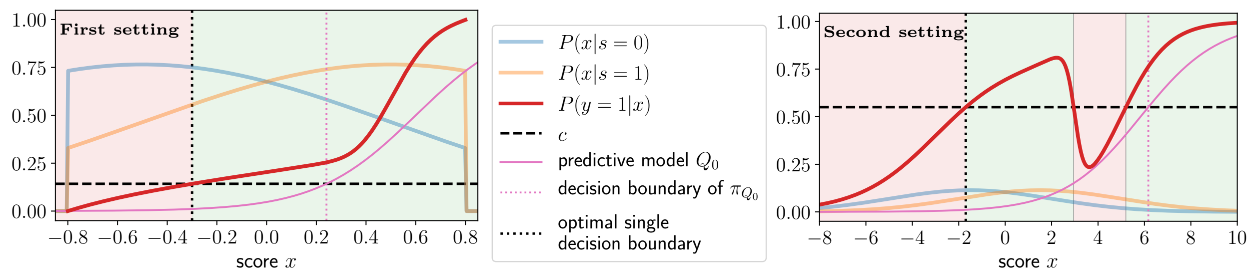

Figure 1: Two synthetic settings. In red, we show $P(y = 1 \,|\, x)$, where the score x is drawn from different distributions for the two groups (blue/orange). For given c (black, dashed), the optimal policy decides $d = 1$ ($d = 0$) in the shaded green (red) regions. The vertical black, dotted line shows the best policy achievable with a single threshold on x. In pink, we show a possible imperfect logistic predictive model and its corresponding (suboptimal) threshold in x.

5 Experiments

In this section, we empirically evaluate our claims via our gradient-based algorithm on synthetic and real-world data. To this end, we learn a sequence of policies $\{\pi_{\boldsymbol{\theta}_t}\}_{t=1}^{T}$ using the following strategies:

- **Optimal**: decisions are taken by the optimal deterministic threshold rule π^* given by eq. (2), i.e., $\pi_t = \pi^*$ for all t. It can only be shown when the ground truth conditional $P(y \,|\, \boldsymbol{x}, s)$ is known.
- **Deterministic**: decisions are taken by deterministic threshold policies $\pi_t = \pi_{Q_t}$, where Q_t are logistic models maximizing label likelihood trained either in an iterative or aggregate sequence.
- **Logistic**: decisions are taken by logistic policies $\pi_t = \pi_{\boldsymbol{\theta}_t}$ trained via Algorithm 1 (Appendix A) either in an iterative or aggregate sequence.
- **Semi-logistic**: decisions are taken by semi-logistic policies $\tilde{\pi}_t = \tilde{\pi}_{\boldsymbol{\theta}_t}$ trained via Algorithm 1 (Appendix A) either in an iterative or aggregate sequence.

It is crucial that while each of the above methods decides over the same set of proposed $\{(\boldsymbol{x}_i, s_i)\}_{i=1}^{N}$ at each time step t, depending on their decisions, they may collect labels for differing subsets and thus receive different amounts of new training data. During learning, we record the following metrics:[5]

- **Utility:** the utility $u_P(\pi_t)$ achieved by the current policy π_t estimated empirically on a held-out dataset, the *test set*, sampled i.i.d. from the ground truth distribution $P(\boldsymbol{x}, s, y)$. This is the utility that the decision maker would obtain if they deployed the current policy π_t at large in the population.
- **Effective utility:** the utility realized during the learning process up to time t, i.e., $\hat{u}(t) = \frac{1}{N \cdot t} \sum_{t' \leq t} \sum_{(\boldsymbol{x}_i, s_i, y_i) \in \mathcal{D}^{t'}} (y_i - c)$, where $\mathcal{D}^{t'}$ are the data in which the policy $\pi_{t'}(d \,|\, \boldsymbol{x}_i, s_i)$ took positive decisions $d_i = 1$ and N is the number of considered examples at each time step t. This is the utility that the decision maker accumulates while learning increasingly better policies.

[5]In all plots, lines are medians and shaded areas are 25 and 75 percentiles over 30 runs.

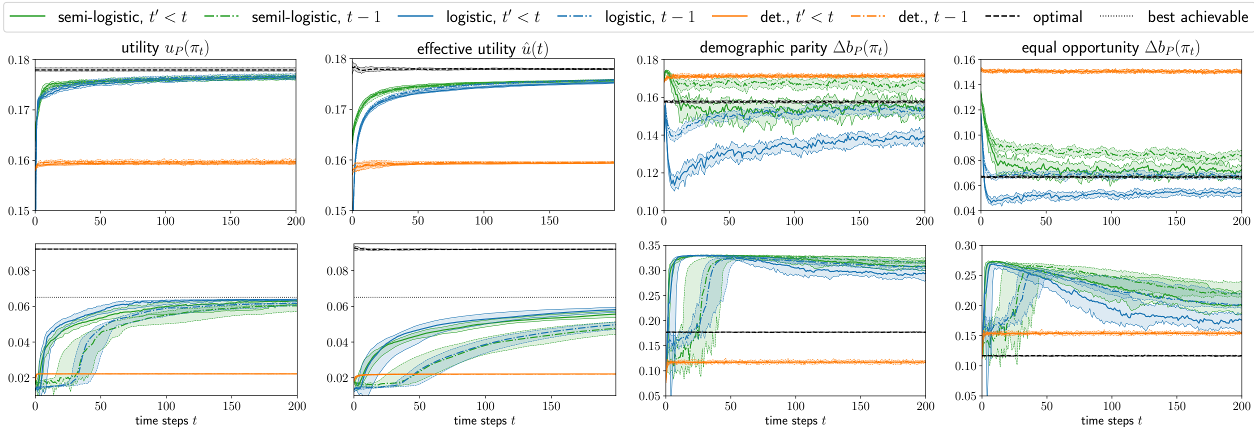

Figure 2: Utility, effective utility, demographic parity and equality of opportunity in synthetic data. The rows correspond to the two settings shown in Figure 1.

- **Fairness:** the difference in group benefits between sensitive groups, i.e., $\Delta b_P(\pi) := b_P^0(\pi) - b_P^1(\pi)$ for both disparate impact ($f(d, y) = d$) and equal opportunity ($f(d, y) = d \cdot y$). A decision policy satisfies the chosen fairness criterion iff $\Delta b_P(\pi) = 0$. Again, this is estimated empirically on the test set and thus measures the level of fairness of π_t that it would achieve in the entire population.

5.1 Experiments on synthetic data

For the ground truth data, we assume that there is only a single non-sensitive feature $x \in \mathbb{R}$ per individual and a sensitive attribute $s \in \{0, 1\}$ indicating group membership. While $P(x \mid s = 0) \neq P(x \mid s = 1)$, in our experiments the policies only take x as input, and *not* the sensitive attribute, which is only used for the fairness constraints. Such single feature scenarios are highly relevant, due to increasing use of score-based decision support systems (e.g., credit scores, or pretrial risk scores), where full training data and the functional form of the score are not available due to, e.g., privacy or intellectual property reasons. For any score that is monotonic in the true probability of a positive outcome there exists a single decision threshold for the score resulting in the optimal policy, which lends additional support to score-based systems.

We consider two different settings, illustrated in Figure 1, where $s \sim \mathrm{Ber}(0.5)$ and the distributions of the fictitious score x differs slightly for the two groups (details in appendix B). In the first setting, the conditional probability $P(y = 1 \mid x)$ is strictly monotonic in the score and does not explicitly depend on s, but is not well calibrated, i.e., $P(y = 1 \mid x)$ is not directly proportional to x. In the second setting, the conditional probability $P(y = 1 \mid x)$ crosses the cost threshold c multiple times, resulting in two disjoint intervals of scores for which the optimal decision is $d = 1$ (green areas). Figure 1 shows the best achievable deterministic threshold rules based on a logistic predictive model.

Figure 2 summarizes the results for $\lambda = 0$, i.e., without fairness constraints. Our method outperforms prediction based deterministic threshold rules in terms of (effective) utility in both experimental settings (rows). This can be easily understood from the evolution of policies illustrated in Figure 4 in

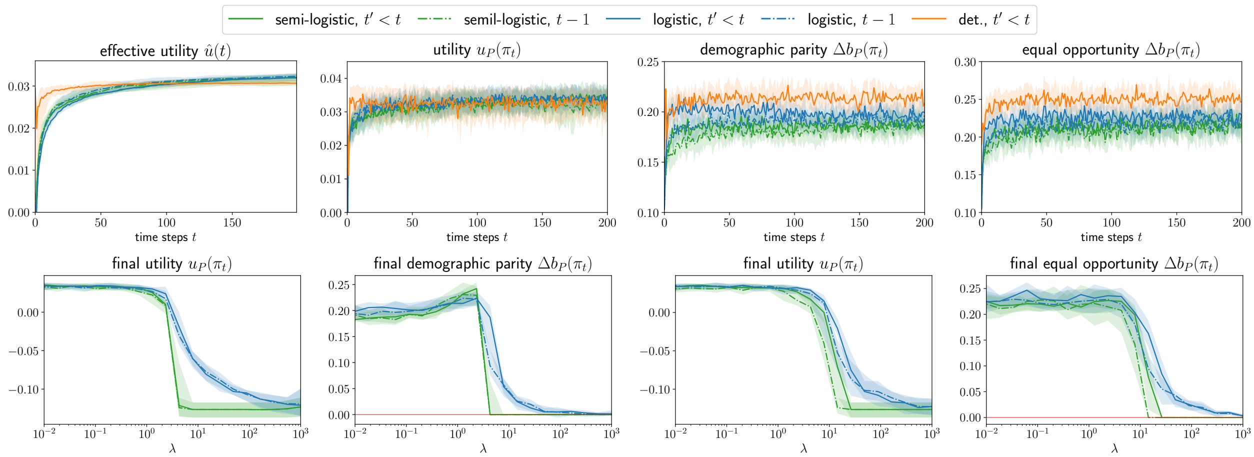

Figure 3: Performance on COMPAS data. The first row shows training progress for $\lambda = 0$, where all four metrics are estimated on the held-out dataset. The second row shows the final ($t = 200$) utility and demographic parity when constraining demographic parity (first and second column), as well as utility and equal opportunity when constraining equal opportunity (third and fourth column) also estimated on the held-out set as a function of λ. The thin red line indicates perfect fairness.

appendix B. In the first setting, the exploring policies locate the optimal decision boundary, whereas the deterministic threshold rules, which are based on learned predictive models, do not, even though $P(y = 1 \,|\, x)$ is monotonic in x and has a sigmoidal shape. In the second setting, our methods explore more and eventually take mostly positive decisions for x right of the vertical dotted line in Figure 1, which is indeed the best achievable single threshold policy. In contrast, non-exploring deterministic threshold rules converge to a suboptimal threshold at $x \approx 5$, ignoring the left green region.

In the first setting, we also observe that the suboptimal predictive models amplify unfairness beyond the levels exhibited by the optimal policy both in terms of demographic parity and equality of opportunity. On the contrary, for our approach levels of unfairness are comparable or even below those of the optimal policy. The second setting shows that depending on the ground truth distribution, higher utility can be directly linked to larger fairness violations. In such cases, our approach allows us to explicitly control for fairness. Results on how utility, demographic parity and equality of opportunity behave when adding fairness constraints with different λ are shown in Figures 5 and 6 in appendix B. In essence, λ continuously trades off (effective) utility and fairness violations up to the point of perfect fairness in the ground truth distribution.

5.2 Experiments on real data

Here, we use the COMPAS recidivism prediction dataset compiled by ProPublica Angwin et al. (2016), which comprises of information about criminal offenders screened through the COMPAS tool in Broward County, Florida during 2013-2014. For each offender, the dataset contains a set of demographic features, the offender's criminal history, and the risk score assigned by COMPAS. Moreover, ProPublica collected whether or not these individuals actually recidivated within two years of the screening. In our experiments, $s \in \{0, 1\}$ indicates whether individuals identify as "white", the label y indicates recidivism, and $d \sim \pi(\boldsymbol{x}, s)$ determines whether an individual is released from jail.

Again, s is not used as an input to the policies. We use 80% of the data to learn the decision policies, where at each step t, we sample (with replacement) N individuals from this set, and the remaining 20% as a held-out set to evaluate each learned policy in the population of interest.

We first summarize the results for $\lambda = 0$, i.e., without explicit fairness constraints in the first row of Figure 3. A slight utility advantage of the deterministic threshold rule in the first time steps is quickly overcome by our exploring policies (second column) with a crossover of effective utility in favor of our strategies roughly around $t = 100$ (first column). Hence, early exploration not only pays off to eventually be able to take better decisions, but also reaps higher profit during training. Moreover, all strategies based on exploring policies consistently achieve lower violations of both fairness metrics than the deterministic threshold rules. In summary, even when not controlling for fairness constraints, i.e., in a pure utility maximization setting, exploring policies achieve higher utility and simultaneously reduce unfairness compared to deterministic threshold rules.

In the second row of Figure 3, we show how utility and demographic parity (equal opportunity) of the final policy $\pi_{t=200}$ change as a function of λ when constraining demographic parity (equal opportunity). As expected, while we are able to achieve perfect demographic parity (equal opportunity), this comes with a drop in utility. All remaining metrics under both constraints are shown in Figure 7 in appendix B. Finally, two remarks are in order. First, for real-world data we cannot evaluate the optimal policy and do not expect it to reside in our model class. However, even when logistic models do not perfectly capture the conditional $P(y = 1 \,|\, \boldsymbol{x})$, our comparisons here are "fair" in that all strategies have equal modeling capacity. Second, here we take the COMPAS dataset as our (empirical) ground truth distribution even though it likely also suffered from selective labels. To learn about the real distribution underlying the dataset, we would need to actually deploy our strategy.

6 Conclusions

In this work, we have analyzed consequential decision making using imperfect predictive models, which are learned from data gathered by potentially biased, historical decisions. First, we have articulated how this approach fails to optimize utility when starting with a faulty deterministic policy. Next, we have presented how directly learning to decide with exploring policies avoids this failure mode while respecting common fairness constraints. Finally, we have introduced and evaluated a simple, yet practical gradient-based algorithm to learn fair exploring policies that improve over time.

Since we have shown how shifting focus from learning predictions to learning decisions requires some form of exploration, we hope to stimulate future research on how to explore ethically and fairly in different domains. Furthermore, we have assumed a static ground truth distribution, which is unaffected by our decisions. Incorporating such feedback or time-varying externalities simultaneously with (then changing) fairness criteria could be of great interest for a variety of applications. Finally, it would be important to evaluate our algorithm using interventional experiments.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1638–1646, Bejing, China, 2014. PMLR.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There is software used across the country to predict future criminals. and it is biased against blacks. *ProPublica, May*, 23, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Athey, S. and Wager, S. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Bottou, L., Peters, J., nonero Candela, J. Q., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Dimitrakakis, C., Liu, Y., Parkes, D., and Radanovic, G. Bayesian fairness. In *AAAI*, 2019.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104. Omnipress, 2011.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226. ACM, 2012.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing. *JMLR*, 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online learning with an unknown fairness metric. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2605–2614. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7526-online-learning-with-an-unknown-fairness-metric.pdf`.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Heidari, H. and Krause, A. Preventing disparate treatment in sequential decision making. In *IJCAI*, pp. 2248–2254, 2018.

Helmbold, D. P., Littlestone, N., and Long, P. M. Apple tasting. *Information and Computation*, 161 (2):85–139, 2000.

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239*, 2018.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *World Wide Web Conference*, WWW '18, pp. 1389–1398, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. *arXiv preprint arXiv:1611.03071*, 2016.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

Jung, J., Shroff, R., Feller, A., and Goel, S. Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868*, 2018.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8909–8920, 2018.

Kallus, N. and Zhou, A. Residual unfairness in fair machine learning from prejudiced data. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2439–2448, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Kiefer, J., Wolfowitz, J., et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2017a.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H. (ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017b. ISBN 978-3-95977-029-3.

Lakkaraju, H. and Rudin, C. Learning Cost-Effective and Interpretable Treatment Regimes. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 166–175, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. 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*, pp. 275–284. ACM, 2017.

Langford, J., Strehl, A., and Wortman, J. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 528–535, New York, NY, USA, 2008. ACM.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3150–3158, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., and Chabris, C. F. Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences*, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1820701116. URL `https://www.pnas.org/content/early/2019/05/08/1820701116`.

Mitchell, S., Potash, E., and Barocas, S. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *FAT*, 2019.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395. JMLR.org, 2014.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 814–823. JMLR.org, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 3231–3239, Cambridge, MA, USA, 2015b. MIT Press.

Tabibian, B., Gomez, V., De, A., Schoelkopf, B., and Gomez-Rodriguez, M. Consequential ranking algorithms and long-term welfare. *arXiv preprint arXiv:1905.05305*, 2019.

Valera, I., Singla, A., and Gomez-Rodriguez, M. Enhancing the accuracy and fairness of human decision making. In *Neural Information Processing Systems*, 2018.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

A Designing exploring policies

In this section, our goal is to put Proposition 4 into practice by designing an algorithm that finds an exploring policy that achieves the *same* utility as the optimal policy π^* using data gathered by a given initial exploring policy π_0, i.e., not from the ground truth distribution $P(\boldsymbol{x}, s, y)$. To this end, we consider a class of parameterized exploring policies $\Pi(\Theta)$ and we aim to find the policy $\pi_{\boldsymbol{\theta}^*} \in \Pi(\Theta)$ that solves the optimization problem in eq. (5).

For a gradient-based approach, note that we can obtain an expression for $\nabla_{\boldsymbol{\theta}_t} v_P(\pi_{\boldsymbol{\theta}_t})$ by simply replacing π_0 with $\pi_{\boldsymbol{\theta}_{t-1}}$ in eq. (8). Thus we can estimate the gradient with samples $(\boldsymbol{x}_i, s_i, y_i)$ from the distribution $P_{\pi_{t-1}}$ induced by the previous policy π_{t-1}, and sample the decisions from the policy under consideration $d_i \sim \pi_{\boldsymbol{\theta}_t}$. This yields an unbiased finite sample Monte-Carlo estimator for the gradients

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_t} u(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) &\approx \frac{1}{n_{t-1}} \sum_{i=1}^{n_{t-1}} \frac{d_i(y_i - c)}{\pi_{\boldsymbol{\theta}_{t-1}}(d = 1 \,|\, \boldsymbol{x}_i, s_i)} \nabla_{\boldsymbol{\theta}_t} \log \pi_{\boldsymbol{\theta}_t}(d_i \,|\, \boldsymbol{x}_i, s_i), \\ \nabla_{\boldsymbol{\theta}_t} b^s(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) &\approx \frac{1}{n_{t-1}} \sum_{i=1}^{n_{t-1}} \frac{f(d_i, y_i)}{\pi_{\boldsymbol{\theta}_{t-1}}(d = 1 \,|\, \boldsymbol{x}_i, s_i)} \nabla_{\boldsymbol{\theta}_t} \log \pi_{\boldsymbol{\theta}_t}(d_i \,|\, \boldsymbol{x}_i, s_i). \end{aligned} \tag{9}$$

where n_{t-1} is the number of positive decisions taken by $\pi_{\boldsymbol{\theta}_{t-1}}$. Here, it is important to notice that, while the decisions by $\pi_{\boldsymbol{\theta}_{t-1}}$ were actually taken and, as a result, (feature and label) data was gathered under $\pi_{\boldsymbol{\theta}_{t-1}}$, the decisions $d_i \sim \pi_{\boldsymbol{\theta}_t}$ are just sampled to implement SGA. The overall policy learning process is summarized in Algorithm 1, where $\text{MINIBATCH}(\mathcal{D}, B)$ samples a minibatch of size B from the dataset $\mathcal{D}$ and $\text{INITIALIZEPOLICY}()$ initializes the policy parameters.

Remarks. In Algorithm 1, to learn each policy π_t, we have limited ourselves to data gathered only by the previous policy π_{t-1}. However, we may readily use samples from the distribution $P_{\pi_{t'}}$ induced by *any* previous policy $\pi_{t'}$ in eq. (9). The average of multiple gradient estimators for several $t' < t$ is again an unbiased gradient estimator. In practice, one may decide to consider recent policies $\pi_{t'}$, which are more similar to π_t, thus ensuring that the gradient estimator does not suffer from high variance.

The way in which we use weighted sampling to estimate the above gradients closely relates to the concept of weighted inverse propensity scoring (wIPS), commonly used in counterfactual learning Bottou et al. (2013); Swaminathan & Joachims (2015a), off-policy reinforcement learning Sutton & Barto (1998), and contextual bandits Langford et al. (2008). However, a key difference is that, in wIPS, the labels y are always observed. Despite this difference, we believe that recent advances to reduce the variance of the gradients in weighted inverse propensity scoring, such as clipped-wIPS Bottou et al. (2013), self-normalized estimator Swaminathan & Joachims (2015b), or doubly robust estimators Dudík et al. (2011), may be also applicable to our setting. This is left for future work.

Logistic policy. Let us now introduce a concrete parameterization of $\pi_{\boldsymbol{\theta}}$, a *logistic policy* given by

$$\pi_{\boldsymbol{\theta}}(d = 1 \,|\, \boldsymbol{x}, s) = \sigma(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}) \in (0, 1),$$

where $\sigma(a) := \frac{1}{1+\exp(-a)}$ is the logistic function, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ are the model parameters, and $\boldsymbol{\phi} : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}^m$ is a fixed feature map. Note that any logistic policy is an exploring policy and

Algorithm 1 ConsequentialLearning: train a sequence of policies $\pi_{\boldsymbol{\theta}_t}$ of increasing $v_P(\pi_{\boldsymbol{\theta}_t})$.

Require: Cost parameter c, number of time steps T, number of decisions N, number of iterations M, minibatch size B, penalty weight λ, and learning rate α.

$\boldsymbol{\theta}_0 \leftarrow$ InitializePolicy()
for $t = 0, \dots, T-1$ **do** ▷ time steps
  $\mathcal{D}^t \leftarrow$ CollectData($\boldsymbol{\theta}_t, N$)
  $\boldsymbol{\theta}_{t+1} \leftarrow$ UpdatePolicy($\boldsymbol{\theta}_t, \mathcal{D}^t, M, B, \alpha$)
return $\{\pi_{\boldsymbol{\theta}_t}\}_{t=0}^{T}$

function CollectData($\boldsymbol{\theta}$, N)
  $\mathcal{D} \leftarrow \emptyset$
  for $i = 1, \dots, N$ **do** ▷ N decisions
    $(\boldsymbol{x}_i, s_i) \sim P(\boldsymbol{x}, s)$
    $d_i \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{x}, s)$
    if $d_i = 1$ **then** ▷ positive decision
      $y_i \sim P(y \mid \boldsymbol{x}, s)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_i, s_i, y_i)\}$
  return $\mathcal{D}$ ▷ data observed under $\pi_{\boldsymbol{\theta}}$

function UpdatePolicy($\boldsymbol{\theta}'$, $\mathcal{D}$, M, B, α)
  $\boldsymbol{\theta}^{(0)} \leftarrow \boldsymbol{\theta}'$
  for $j = 1, \dots, M$ **do** ▷ iterations
    $\mathcal{D}^{(j)} \leftarrow$ Minibatch($\mathcal{D}, B$) ▷ sample minibatch
    $\nabla \leftarrow 0$, $n_j \leftarrow 0$
    for $(\boldsymbol{x}, s, y) \in \mathcal{D}^{(j)}$ **do** ▷ accumulate gradients for current mini batch
      $d \sim \pi_{\boldsymbol{\theta}^{(j)}}(\boldsymbol{x}, s)$
      if $d = 1$ **then**
        $n_j \leftarrow n_j + 1$
        $\nabla \leftarrow \nabla + \nabla_{\boldsymbol{\theta}} u(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}$
          $+\lambda \left(b^0(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) - b^1(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})\right) \left(\nabla_{\boldsymbol{\theta}} b^0(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}} - \nabla_{\boldsymbol{\theta}} b^1(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(j)}}\right)$
    $\boldsymbol{\theta}^{(j+1)} \leftarrow \boldsymbol{\theta}^{(j)} + \alpha \frac{\nabla}{n_j}$
  return $\boldsymbol{\theta}^M$

we can analytically compute its score function $\nabla_{\boldsymbol{\theta}_t} \log \pi_{\boldsymbol{\theta}_t}(d = 1 \mid \boldsymbol{x}, s)$ as

$$\nabla_{\boldsymbol{\theta}_t} \log(\sigma(\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t)) = \frac{\boldsymbol{\phi}_i}{1 + e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t}} \in \mathbb{R}^m,$$

where $\boldsymbol{\phi}_i := \boldsymbol{\phi}(\boldsymbol{x}_i, s_i)$. Using this expression, we can rewrite the empirical estimator for the gradient in eq. (9)

$$\nabla_{\boldsymbol{\theta}_t} u(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) \approx \frac{1}{n_{t-1}} \sum_{i=1}^{n_{t-1}} \frac{1 + e^{-\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1}}}{1 + e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t}} \, d_i \, (y_i - c) \, \boldsymbol{\phi}_i,$$
$$\nabla_{\boldsymbol{\theta}_t} b^s(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) \approx \frac{1}{n_{t-1}} \sum_{i=1}^{n_{t-1}} \frac{1 + e^{-\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1}}}{1 + e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t}} \, f(d_i, y_i) \, \boldsymbol{\phi}_i.$$

Given the above expression, we have all the necessary ingredients to implement Algorithm 1.

Semi-logistic policy. As discussed in the previous section, randomizing decisions may be questionable in certain practical scenarios. For example, in loan decisions, it may appear wasteful for the bank and contestable for the applicant to deny a loan with probability greater than zero to individuals who are believed to repay by the current model. In those cases, one may consider the following modification of the logistic policy, which we refer to as *semi-logistic policy*:

$$\tilde{\pi}_{\boldsymbol{\theta}}(d = 1 \,|\, \boldsymbol{x}, s) = \begin{cases} 1 & \text{if } \boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} \geq 0, \\ \sigma(\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}) & \text{if } \boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} < 0. \end{cases}$$

Similarly as in the logistic policy, we can compute the score function analytically as:

$$\nabla_{\boldsymbol{\theta}} \log \tilde{\pi}_{\boldsymbol{\theta}}(d \,|\, \boldsymbol{x}, s) = \frac{\boldsymbol{\phi}(\boldsymbol{x}, s)}{1 + e^{\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta}}} \, \mathbf{1}[\boldsymbol{\phi}(\boldsymbol{x}, s)^\top \boldsymbol{\theta} < 0],$$

and use this expression to compute an unbiased estimator for the gradient in eq. (9) as:

$$\nabla_{\boldsymbol{\theta}_t} u(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) \approx \frac{1}{n_{t-1}} \sum_{\substack{i=1 \\ \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t < 0}}^{n_{t-1}} \frac{d_i \, (y_i - c) \, \boldsymbol{\phi}_i}{1 + e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t}} \times \begin{cases} 1 & \text{if } \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1} \geq 0, \\ (1 + e^{-\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1}}) & \text{if } \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1} < 0. \end{cases}$$

$$\nabla_{\boldsymbol{\theta}_t} b^s(\pi_{\boldsymbol{\theta}_t}, \pi_{\boldsymbol{\theta}_{t-1}}) \approx \frac{1}{n_{t-1}} \sum_{\substack{i=1 \\ \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t < 0}}^{n_{t-1}} \frac{f(d_i, y_i) \, \boldsymbol{\phi}_i}{1 + e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_t}} \times \begin{cases} 1 & \text{if } \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1} \geq 0, \\ (1 + e^{-\boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1}}) & \text{if } \boldsymbol{\phi}_i^\top \boldsymbol{\theta}_{t-1} < 0. \end{cases}$$

Note that the semi-logistic policy is an exploring policy and thus satisfies the assumptions of Proposition 4.

Finally, in all our experiments, we directly worked with the available features $\boldsymbol{x}$ as inputs and added a constant offset, i.e., $\phi(\boldsymbol{x}, s) = (1, \boldsymbol{x})$.

B Additional experimental results

B.1 Experiments on synthetic data

Setup. The precise setup for the two different synthetic settings, illustrated in Figure 1, is as follows. The only feature x is a scalar score and $s \sim \text{Ber}(0.5)$. In the first setting, x is sampled from a normal distribution $\mathcal{N}(\mu = s - 0.5, \sigma = 3.5)$ truncated to $x \in [-0.8, 0.8]$, and the conditional probability $P(y \,|\, x)$ is strictly monotonic in the score and does not explicitly depend on s. As a result, for any c, there exists a single decision boundary for the score that results in the optimal policy, which is contained in the class of logistic policies. Note, however, that the score is not well calibrated, i.e., $P(y \,|\, x)$ is not directly proportional to x.

In the second setting, $x \sim \mathcal{N}(\mu = 3(s - 0.5), \sigma = 3.5)$. Here, the conditional probability $P(y \,|\, x)$ crosses the cost threshold c multiple times, resulting in two disjoint intervals of scores for which the optimal decision is $d = 1$ (green areas). Consequently, the optimal policy cannot be implemented by a deterministic threshold rule based on a logistic predictive model. We show the best achievable single decision threshold in Figure 1.

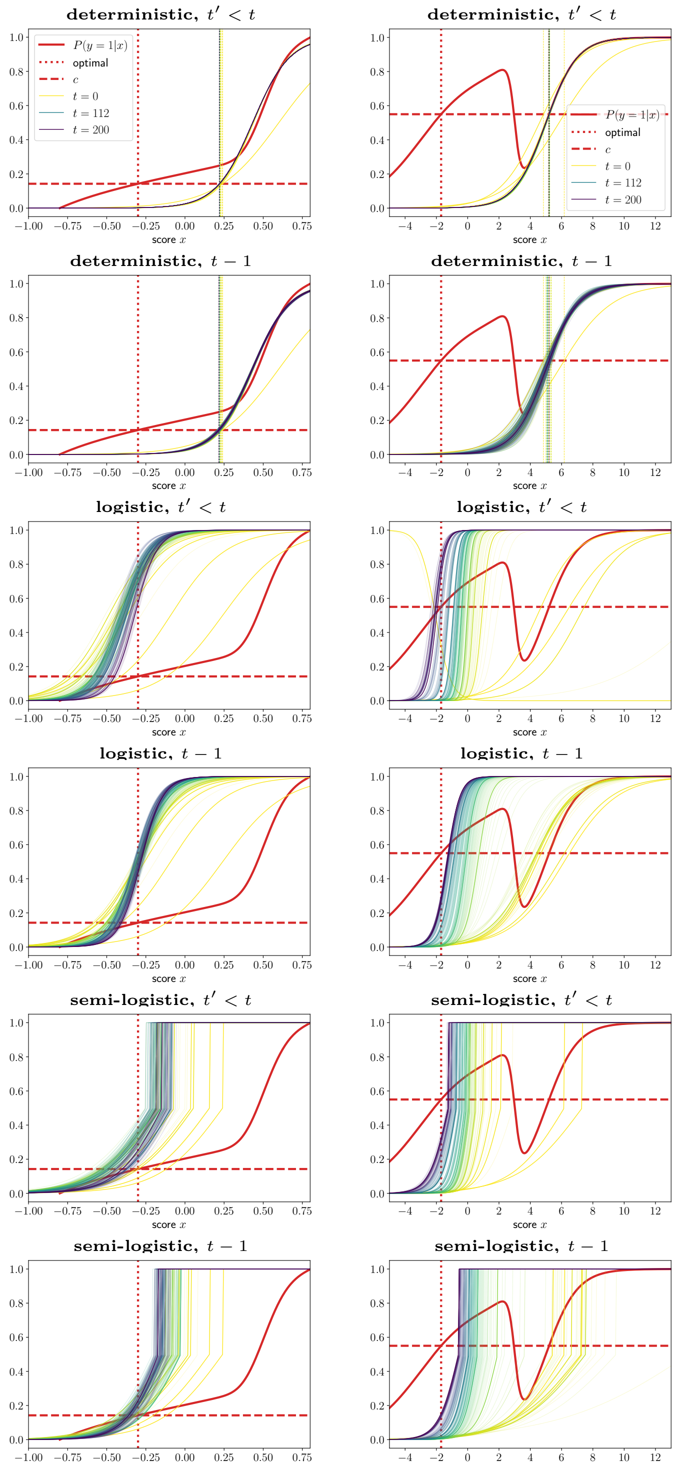

Figure 4: Learned predictive models for deterministic threshold rules and learned policies for the (semi-)logistic policies. The columns correspond to the two synthetic settings. We overlay the ground truth distribution $P(y = 1 \mid x)$ (red line), cost parameter c (dashed, red), and optimal single decision boundary in x within our model class (dotted, red). We describe the plots in detail in the text.

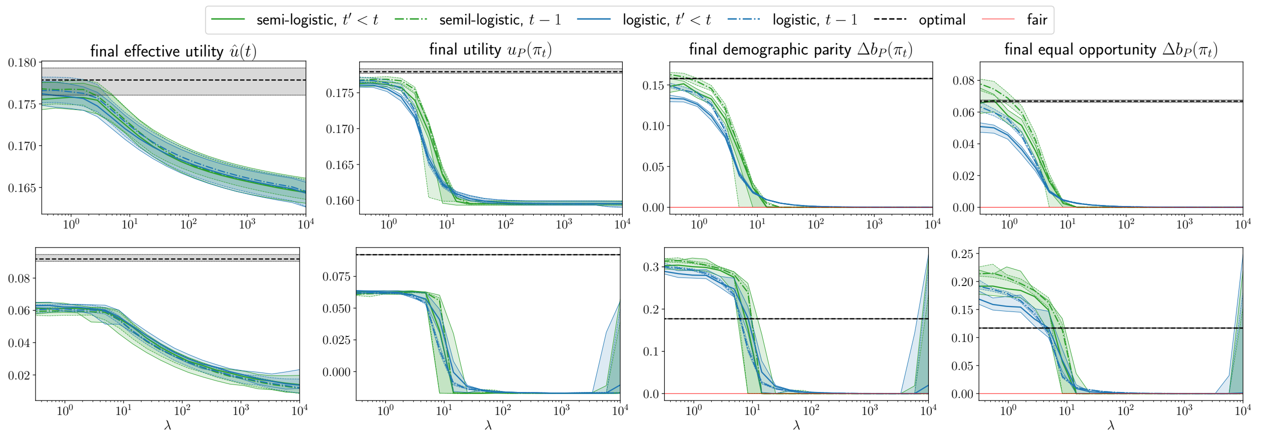

Figure 5: We show (effective) utility, effective utility, demographic parity, and equal opportunity (columns) at the final time step $t = 200$ as a function of λ where we constrain demographic parity, i.e., $f(d, y) = d$. The first row corresponds to the first setting and the second row corresponds to the second setting.

Evolution of policies. In Figure 4 we show for a representative run at $\lambda = 0$ how the different policies evolve in the two synthetic settings over time. The two columns correspond to the two different synthetic settings. For all policies, we show snapshots at a fixed number of logarithmically spaced time steps between $t = 0$ and $t = 200$. For deterministic threshold rules, we show the logistic function of the underlying predictive model. The vertical dashed line corresponds to the decision boundary in x. For the logistic and semi-logistic policies, the lines correspond to $\pi_t(d = 1 \mid x)$, i.e., to the probability of giving a positive decision for a given input x. Note that the semi-logistic policies have a discontinuity, because we do not randomize for which the model believes $d = 1$ is a favorable decision. For reference, we also show the true conditional distribution, the cost parameter as well as the best achievable single decision boundary.

In the first setting, the exploring policies locate the optimal decision boundary, whereas the deterministic threshold rules, which are based on learned predictive models, do not, even though $P(y = 1 \mid x)$ is monotonic in x and has a sigmoidal shape. The predictive models focus on fit the rightmost part of the conditional well, but ignore the right region, from which they never receive data.

In the second setting, our methods explore more and eventually take mostly positive decisions for x right of the vertical dotted line in Figure 1, which is indeed the best achievable single threshold policy. In contrast, non-exploring deterministic threshold rules again suffer from the same issue as in the first setting and converge to a suboptimal threshold at $x \approx 5$. They ignoring the left green region in Figure 1 and do not overcome the dip of $P(y = 1 \mid x)$ below c, because they never receive data for $x \leq 4$.

Adding fairness constraints. Figures 5 and 5 show how all four metrics at the final time step $t = 200$ evolve as λ is increased over the range $[10^{-0.5}, 10^4]$. In Figure 5 we use demographic parity in the fairness constraint, i.e., $f(d, y) = d$, whereas in Figure 6 we use equal opportunity as a fairness constraint, i.e., $f(d, y) = d \cdot y$. In both figures, the first row corresponds to the first setting and the second row corresponds to the second setting. In both cases, our approach achieves perfect fairness for sufficiently large λ at the expected cost of a drop in (effective) utility. Interestingly, in the two

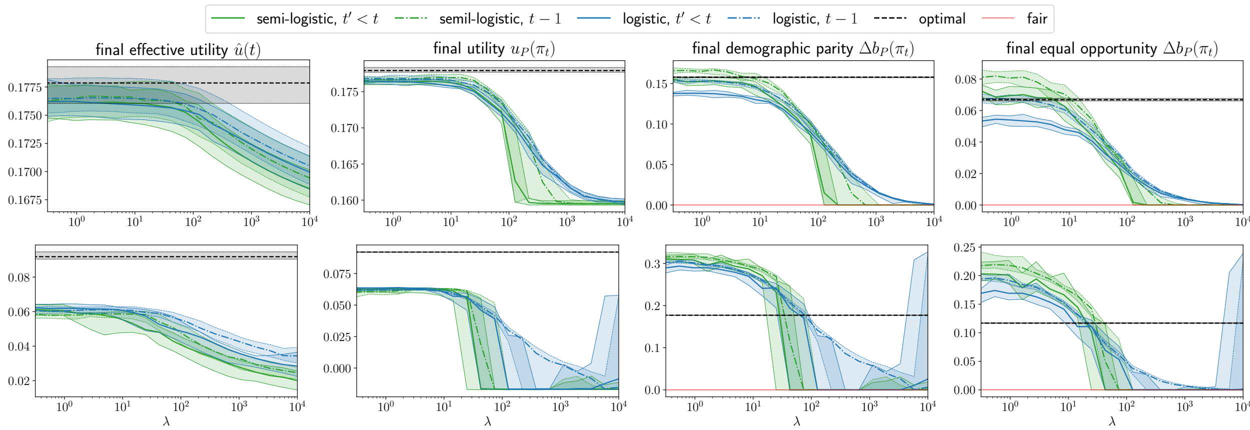

Figure 6: We show (effective) utility, demographic parity, and equal opportunity (columns) at the final time step $t = 200$ as a function of λ where we constrain equal opportunity, i.e., $f(d, y) = d \cdot y$. The first row corresponds to the first setting and the second row corresponds to the second setting.

selected synthetic settings, enforcing demographic parity, also leads to satisfying equal opportunity, and—to a lesser extent—also vice versa.

B.2 Experiments on real data

Analogously to Figures 5 and 6, we show the effect of enforcing fairness constraints in the COMPAS dataset in Figure 7. Here, the first row corresponds to using demographic parity as a fairness measure, while the second row corresponds to using equal opportunity as a fairness measure. The overall trends are similar to the results we have observed in the synthetic settings, reinforcing the applicability of our approach on real-world data.

B.3 Parameter settings

The parameters used for the different experiments have been found by few iterations of trial. The number of time steps is $T = 200$ for all datasets. For the first synthetic setting we used $\alpha = 1$, $B = 256$, $M = 128$, $N = B \cdot M$, and $c \approx 0.142$ (chosen such that the optimal decision boundary is at $x = -0.3$). For the second synthetic setting we used $\alpha = 0.5$, $B = 512$, $M = 32$, $N = B \cdot M$, and $c = 0.55$. Here we also decay the learning rate by a factor of 0.8 every 30 time steps. For the COMPAS dataset we used $\alpha = 0.1$, $B = 64$, $M = 40 \cdot B$, $N = B^2$, and $c = 0.6$. While the initialization for the synthetic settings can be seen in Figure 4, for COMPAS we trained a logistic predictive model on 500 i.i.d. examples for initializing policies and predictive models.

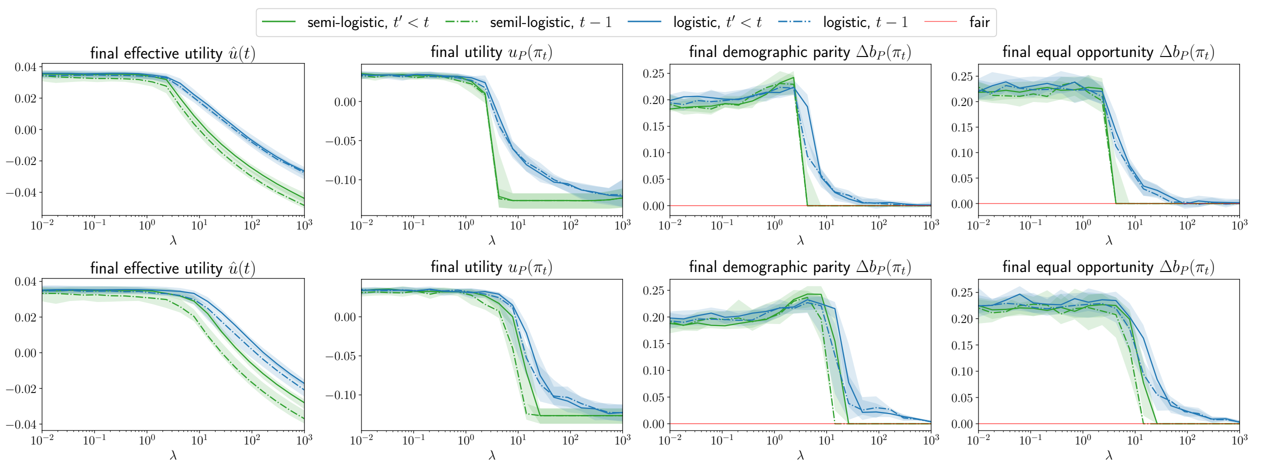

Figure 7: We show (effective) utility, demographic parity, and equal opportunity (columns) for the COMPAS dataset at the final time step $t = 200$ estimated on the held-out dataset as a function of λ. In the first row, we constrain demographic parity, i.e., $f(d, y) = d$, and in the second row we constrain equal opportunity, i.e., $f(d, y) = d \cdot y$.