A PROOFS

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 $v_P(\pi_{Q_0^*}) < v_P(\pi_{Q^*})$. □

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

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

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. □

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 $t \in \mathbb{N}$. □

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

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

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

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

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

$$\begin{aligned} \lim_{n \to \infty} v_{P_{\pi_0}}(\pi_n, \pi_0) = \lim_{n \to \infty} \Big(& u_{P_{\pi_0}}(\pi_n, \pi_0) \\ & - \frac{\lambda}{2} \big(b^0_{P_{\pi_0}}(\pi_n, \pi_0) - b^1_{P_{\pi_0}}(\pi_n, \pi_0) \big)^2 \Big) \end{aligned}$$

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

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

Similarly, for the benefit terms with $f(d, y) = d$ or

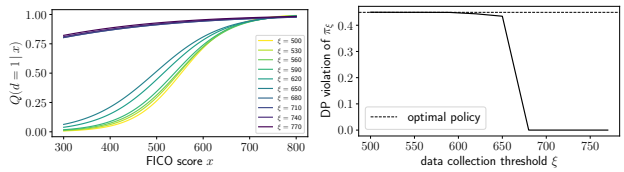

Figure 6: We show the predictive models Q_ξ learned from data collected with an initial threshold of ξ (left) and their violation of demographic parity (right).

$f(d, y) = d \cdot y$ 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

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

□

B DETAILS OF THE LENDING EXAMPLE

For a range of initial data collection score thresholds $\xi \in [500, 800]$, we sample 10,000 scores from the specified population (80% white, 20% black) via inverse transform sampling given the cumulative distributions functions over scores of the two groups. The relatively large number of examples is chosen to illustrate that the negative result is not a consequence of insufficient data. We then fit an L2 regularized logistic regression model to each of these datasets using 5-fold cross validation to select the regularization parameter. This results in a predictive model Q_ξ for each initial data collection threshold ξ. For each of these models we construct the decision rule $\pi_\xi(d = 1\,|\,x) = \mathbf{1}[Q_\xi(y = 1\,|\,x) > c]$, with $c = 0.7$. We then estimate utility and fairness violation of both equal opportunity as well as demographic parity on a large sample from the entire population (one million examples). For completeness, Figure 6 shows the resulting logistic models as well as the violation of demographic parity.

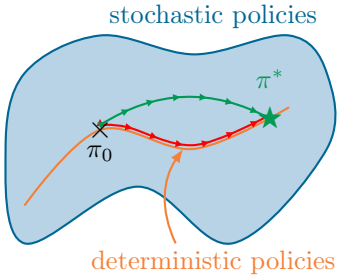

Figure 7: This figure illustrates how it can be impossible to find the optimal policy when the allowed set of policies is restricted to deterministic decision rules.

C ILLUSTRATION OF IMPOSSIBILITY RESULT

Figure 7 illustrates that, even though the optimal policy π^* is deterministic, when starting from a deterministic initial policy π_0, we cannot iteratively reach π^* when updating solely within deterministic policies (red line). It is necessary to deploy stochastic exploring policies along the way to then be able to converge to the optimal policy (green line).

D 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

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

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 MINIBATCH$(\mathcal{D}, B)$ samples a minibatch of size B from the dataset $\mathcal{D}$ and 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. As an example, in the case of counterfactual learning one may interpret $\pi_0(x, s)$ in Eq. 4 as a treatment assignment mechanism in a randomized control trial. Under this interpretation, the two most prominent differences with respect to the literature become apparent. First, we do not observe outcomes in the control group. Second, in observational studies for treatment effect estimation (Rubin, 2005), one usually estimates the direct causal effect of d on y, i.e., $P(y|do(d = d'), x, s)$, in the presence of confounders x, s that affect both d and y. This could be evaluated in a (partially) randomized control trial, where IPW also comes in naturally (Pearl, 2009). In contrast, in our setting, the true label y is independent of the decision d and we estimate the conditional $P(y|x, s)$ using data from the induced distribution $P_{\pi_0}(x, s) \propto P(x, s)\pi_0(x, s)$. With exploring policies, we obtain indirect access to the true data distribution $P(x, s)$ (positivity), and thus to an unbias estimator of the conditional distribution $P(y|x, s)$ (consistency).

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

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 Experiments on Synthetic Data

Setup. The precise setup for the two different synthetic settings, illustrated in Figure 2, 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 $\mathcal{N}(\mu = 0.5 - s, \sigma = 1)$ 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, $x \sim \mathcal{N}(\mu = 3(0.5 - s), \sigma = 3.5)$. 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 2.

Repeated figure. First, in Figure 8 we again show the contents of Figure 3 in the main text, but added effective utility and also show shaded regions for the 25th and 75th percentile over 30 runs.

Evolution of policies. In Figure 9 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 $\pi_t(d = 1 \,|\, x)$, 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 $P(y = 1 \,|\, x)$ 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 2, 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 2 and do not overcome the dip of $P(y = 1 \,|\, x)$ below c, because they never receive data for $x \leq 4$.

Adding fairness constraints. Figures 10 and 10 show how all four metrics at the final time step $t = 200$ evolve as λ is increased. In Figure 10 we use demographic parity in the fairness constraint, i.e., $f(d, y) = d$, whereas in Figure 11 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 reduced fairness violations for sufficiently large λ at the expected cost of a drop in (effective) utility. Interestingly, in the two selected synthetic settings, enforcing demographic parity, also leads to satisfying equal opportunity, and—to a lesser extent—also vice versa.

E.2 Experiments on Real Data

First, in Figure 12 we again show the contents of Figures 4 and 5 in the main text with shaded regions for the 25th and 75th percentile over 30 runs.

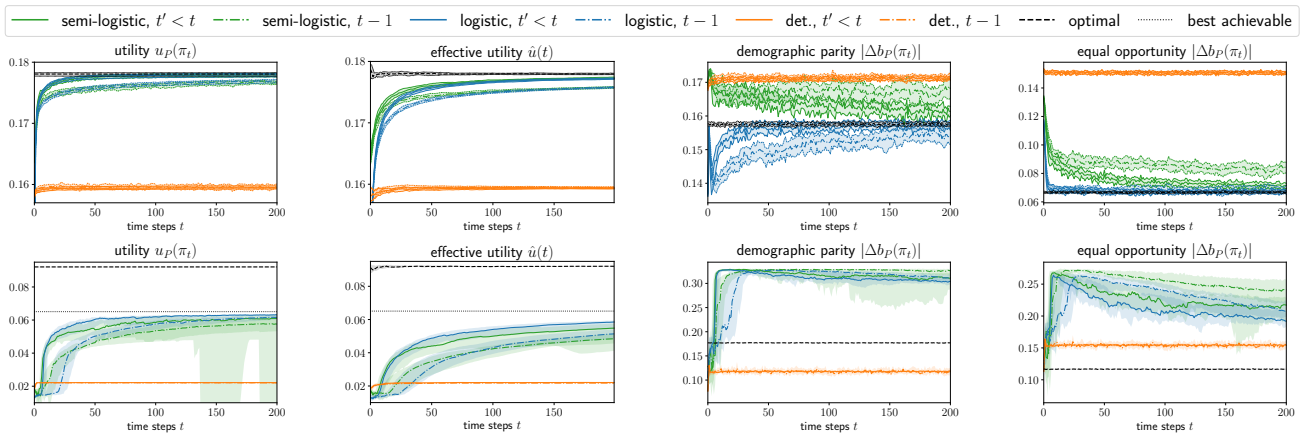

Figure 8: Utility, effective utility, demographic parity and equality of opportunity in the synthetic settings of Figure 2.

Analogously to Figures 10 and 11, we show the effect of enforcing fairness constraints in the COMPAS dataset in Figure 13. 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.

E.3 Parameter Settings

The parameters used for the different experiments have been found by few manual trials. 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 $\alpha = 0.5$, $B = 128$, $M = 32$, $N = B \cdot M$, 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 9, for COMPAS we trained a logistic predictive model on 500 i.i.d. examples for initializing policies and predictive models. For the strategies where we use data from all previous policies, we subsample uniformly at random to always keep the number of gradient updates constant. We also cap the buffer for previously collected data at a maximum size of 10^6, continuously removing the oldest examples as new data is collected.

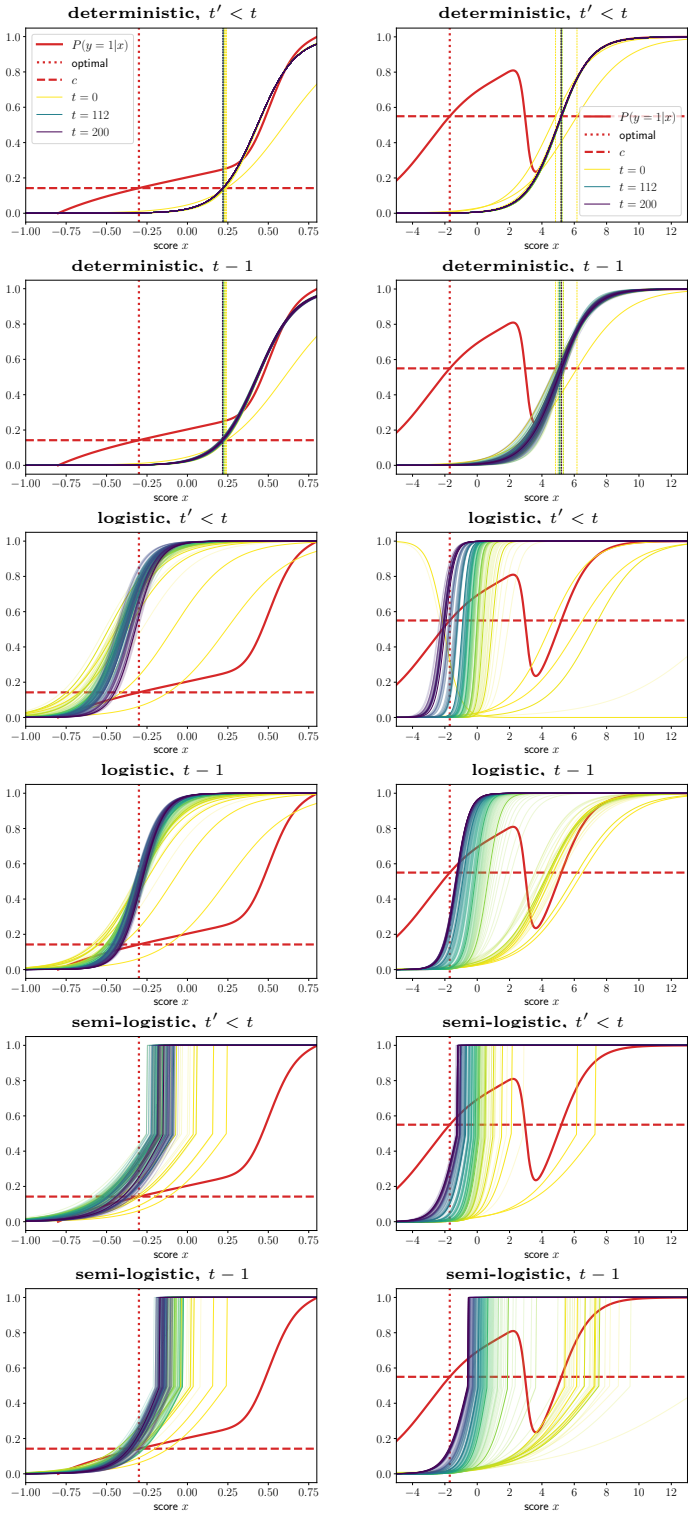

Figure 9: Learned predictive models for deterministic threshold rules and learned policies for the (semi-)logistic policies. The columns correspond to the two synthetic settings. We overlay the ground truth distribution $P(y = 1 \,|\, x)$ (red line), cost parameter c (dashed, red), and optimal single decision boundary in x within our model class (dotted, red). We describe the plots in detail in the text.

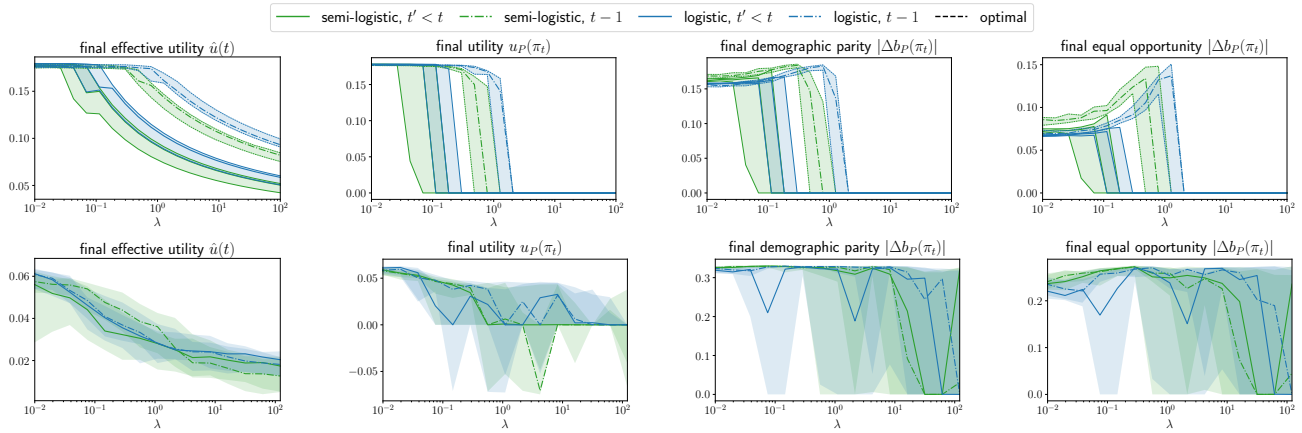

Figure 10: We show (effective) utility, effective utility, demographic parity, and equal opportunity (columns) at the final time step $t = 200$ as a function of λ where we constrain demographic parity, i.e., $f(d, y) = d$. The first row corresponds to the first setting and the second row corresponds to the second setting.

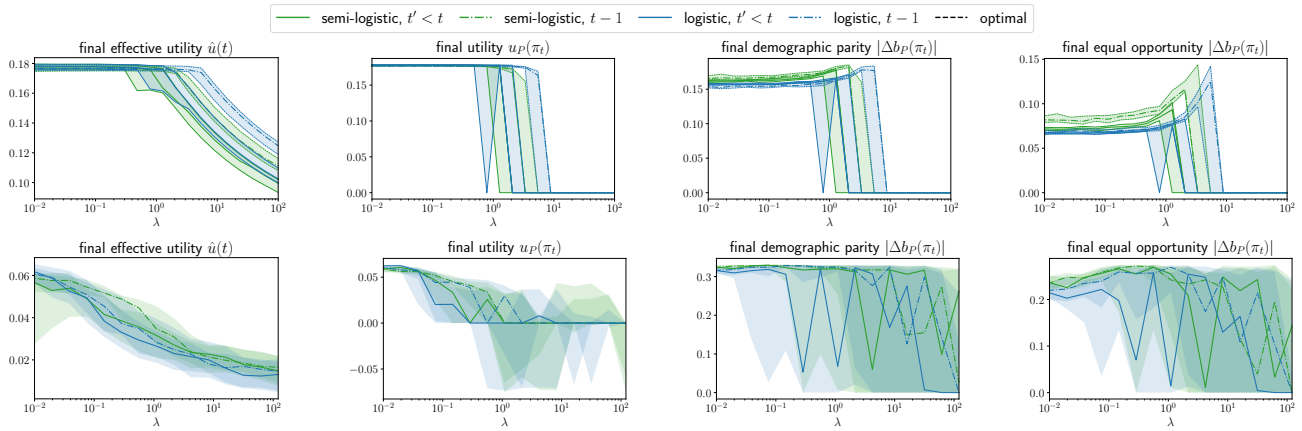

Figure 11: We show (effective) utility, demographic parity, and equal opportunity (columns) at the final time step $t = 200$ as a function of λ where we constrain equal opportunity, i.e., $f(d, y) = d \cdot y$. The first row corresponds to the first setting and the second row corresponds to the second setting.

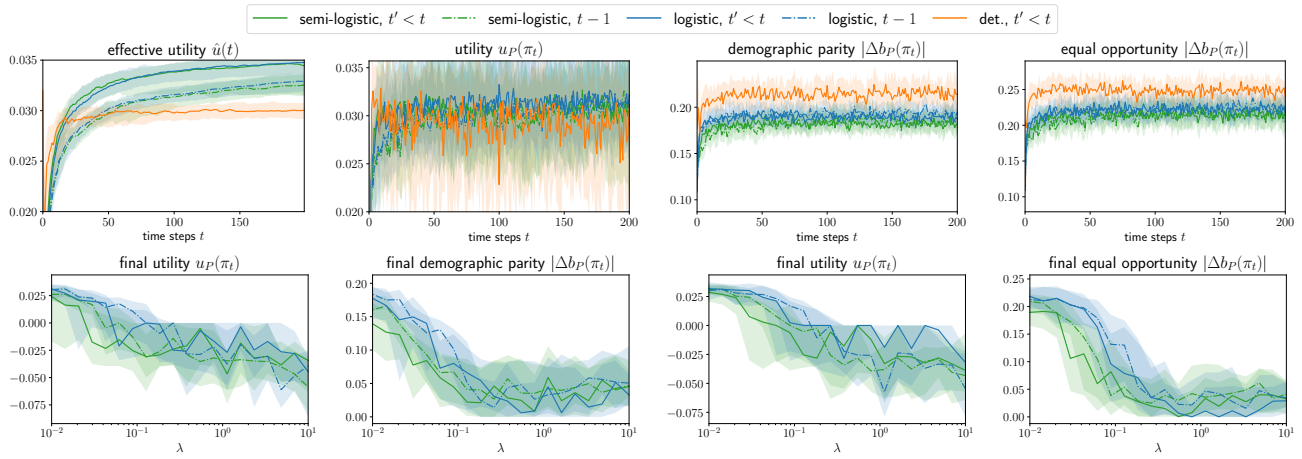

Figure 12: Performance on COMPAS data. The first row shows training progress for $\lambda = 0$, where all four metrics are estimated on the held-out dataset. The second row shows the final ($t = 200$) utility and demographic parity when constraining demographic parity (first and second column), as well 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.

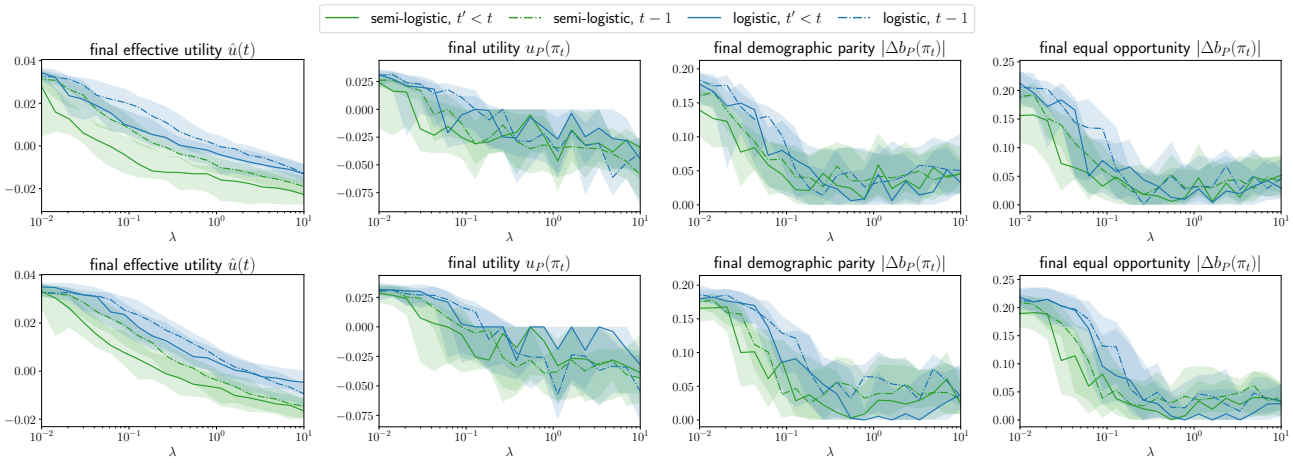

Figure 13: We show (effective) utility, demographic parity, and equal opportunity (columns) for the COMPAS dataset at the final time step $t = 200$ estimated on the held-out dataset as a function of λ. In the first row, we constrain demographic parity, i.e., $f(d, y) = d$, and in the second row we constrain equal opportunity, i.e., $f(d, y) = d \cdot y$.