A Batch Decisions

We describe how our algorithm can be extended to the setting where the decision-maker makes decisions about batches of individuals jointly rather than one individual at a time. For instance, a bank might decide on a portfolio of loans to target at once rather than decide independently for each individual. The challenge is ensuring decisions are fair not just across different batches of individuals, but also across individuals within a batch, since these decisions may be correlated.

In this setting, the state space becomes $S' = S^k = (Z \times \tilde{S})^k$, where there are k individuals and S is the state space of individual i. The action space is $A' = A^k = \{0,1\}^k$—i.e., a binary decision for each individual. Then, the state-action distribution is $\lambda \in \mathbb{R}^{|S'|\times|A'|}$, with a component $\lambda_{s,a} = \lambda_{(s_1,...,s_k),(a_1,...,a_k)}$ for each state $s = (s_1, ..., s_k) \in S'$ and action $a = (a_1, ..., a_k) \in A'$. The policy π can simultaneously make decisions for all k individuals. We let the individual rewards for individual i be $\rho^i \in \mathbb{R}^{|S'|\times|A'|}$. Then, the natural generalization of our fairness constraint is that decisions should be fair on average across both the initial state distribution and across individuals in a single batch. For instance, demographic parity says that

$$\left|\frac{1}{k}\sum_{i=1}^{k}\rho_{i,\mathrm{maj}}^{(\pi)} - \frac{1}{k}\sum_{i=1}^{k}\rho_{i,\mathrm{min}}^{(\pi)}\right| \leq \epsilon,$$

where

$$\begin{aligned}\rho_{i,z}^{(\pi)} &= \mathbb{E}_{(s,a)\sim\Lambda_{i,z}^{(\pi)}}[\rho_{s,a}^{i}]\\ \Lambda_{i,z}^{(\pi)} &= \Lambda^{(\pi)} \mid \exists \tilde{s}_i \in \tilde{S} \;.\; s_0 = (..., (z, \tilde{s}_i), ...),\end{aligned}$$

This constraint can be encoded in our linear program in Algorithm 1 by replacing the second constraint with the following (the objective and first constraint remain the same, except with S replaced by S' and A replaced by A'):

$$|X_{\mathrm{maj}} - X_{\mathrm{min}}| \leq \epsilon$$

where

$$X_z =$$

$$\frac{1}{k}\sum_{i=1}^{k} p_{i,z}^{-1} \sum_{s_j\in S_j : j\neq i}\;\sum_{\tilde{s}_i\in\tilde{S}}\;\sum_{a\in A'} \lambda_{(\ldots,(z,\tilde{s}_i),\ldots),a}\,\rho_{(\ldots,(z,\tilde{s}_i),\ldots),a},$$

and where $p_{i,z}$ is a normalizing constant similar to p_z in the original constraint. Intuitively, this constraint is the same as the original one except that we marginalize over all other individuals (i.e., the sums over s_j for $j \neq i$), and then we average over individuals i as in our fairness constraint. This constraint can similarly be incorporated into Algorithm 2. Finally, while this MDP has number of states exponential in the number of individuals k, this blowup is inevitable since the policy is allowed to make complex decisions based on the states of all individuals.

B Proof of Theorem 2.5

For the first claim, consider the MDP M. The states are $s_0, s_1, s_2, s_3, s_4 \in \tilde{S} \times Z$, where:

$$\begin{aligned}s_0 &= (0, \mathrm{maj})\\ s_1 &= (1, \mathrm{maj})\\ s_2 &= (0, \mathrm{min})\\ s_3 &= (1, \mathrm{min})\\ s_4 &= (2, \mathrm{min}).\end{aligned}$$

The actions are $A = \{0, 1\}$. The transitions are

$$\begin{aligned}P_{s_0,a,s_1} &= 1\\ P_{s_1,a,s_1} &= 1\\ P_{s_2,a,s_3} &= \mathbb{I}[a = 0]\\ P_{s_2,a,s_4} &= \mathbb{I}[a = 1]\\ P_{s_3,s_3} &= 1\\ P_{s_4,s_4} &= 1\end{aligned}$$

for all $a \in A$. The initial distribution is

$$\begin{aligned}D_{s_0} = D_{s_2} &= \frac{1}{2}\\ D_{s_1} = D_{s_3} = D_{s_4} &= 0.\end{aligned}$$

The discount factor is $\gamma = \frac{1}{2}$. The individual rewards are

$$\begin{aligned}\rho_{s_0,a} &= 0\\ \rho_{s_1,a} &= 1\\ \rho_{s_2,a} &= 0\\ \rho_{s_3,a} &= 0\\ \rho_{s_4,a} &= 2,\end{aligned}$$

for all $a \in A$. Let $\pi : S \to A$ be a deterministic policy. It is clear that the only value of π that matters is $\pi(s_2)$. Conditioned on $z = \mathrm{maj}$, regardless of π, the expected cumulative individual reward is

$$\begin{aligned}\mathbb{E}_{(s,a)\sim\Lambda_{\mathrm{maj}}^{(\pi)}}[\rho_{s,a}] &= \left(1 - \frac{1}{2}\right)\sum_{t=1}^{\infty}\frac{1}{2^t}\\ &= \frac{1}{2}.\end{aligned}$$

Conditioned on $z = \mathrm{min}$, if $\pi(s_2) = 0$, then

$$\mathbb{E}_{(s,a)\sim\Lambda_{\mathrm{min}}^{(\pi)}}[\rho_{s,a}] = \begin{cases}0 & \text{if } \pi(s_2) = 0\\ 1 & \text{if } \pi(s_2) = 1.\end{cases}$$

Thus, for $\epsilon < \frac{1}{2}$, it is impossible for the demographic parity constraint to be satisfied.

However, consider the stochastic policy

$$\pi_{s_2,0} = \pi_{s_2,1} = \frac{1}{2}.$$

Then,

$$\mathbb{E}_{(s,a)\sim\Lambda_{\text{min}}^{(\pi)}}[\rho_{s,a}] = \frac{1}{2},$$

so this policy satisfies the demographic parity constraint.

For the second claim, consider the same MDP, except where

$$\rho_{s_4,a} = 0$$

for all $a \in A$. Then, it is clear that

$$\mathbb{E}_{(s,a)\sim\Lambda_{\text{min}}^{(\pi)}}[\rho_{s,a}] = 0$$

regardless of π. Thus, for $\epsilon < \frac{1}{2}$, the demographic parity constraint cannot be satisfied—i.e., $\Pi_{\text{DP},\epsilon} = \varnothing$. □

C Proof of Theorem 3.1

Our proof proceeds in three steps. First, we show that any feasible point of the LP in Algorithm 1 is the state-action distribution $\Lambda^{(\pi)}$ for some policy $\pi \in \Pi_{\text{DP}}$. Second, we show that conversely, for any fair policy $\pi \in \Pi_{\text{DP}}$, the state-action distribution $\Lambda^{(\pi)}$ is a feasible point of the LP. Finally, we combine these two results to prove the theorem.

Step 1. Let $\pi \in \Pi_{\text{DP}}$ be any policy satisfying demographic parity. Then, we claim that the state-action distribution $\Lambda^{(\pi)}$ is a feasible point of the LP in Algorithm 1.

First, we show that $\Lambda^{(\pi)}$ satisfies the first constraint

$$\sum_{a\in A}\Lambda_{s',a}^{(\pi)} = (1-\gamma)D_{s'} + \gamma\sum_{s\in S}\sum_{a\in A}\Lambda_{s,a}^{(\pi)}P_{s,a,s'}$$

for each $s' \in S$.

To this end, note that by induction,

$$D^{(\pi,t)} = (P^{(\pi)})^t D,$$

so

$$D^{(\pi)} = (1-\gamma)\left[\sum_{t=0}^{\infty}(\gamma P^{(\pi)})^t\right]D. \tag{9}$$

Multiplying each side of (9) by $I - \gamma P^{(\pi)}$ (where I is the $|S| \times |S|$ identity matrix), we have

$$\begin{aligned}&(I-\gamma P^{(\pi)})D^{(\pi)}\\ &= (1-\gamma)\left[\sum_{t=0}^{\infty}(\gamma P^{(\pi)})^t - \sum_{t=1}^{\infty}(\gamma P^{(\pi)})^t\right]D\\ &= (1-\gamma)\cdot D.\end{aligned}$$

Note that these algebraic manipulations are valid since the eigenvalues of $\gamma P^{(\pi)}$ are bounded in norm by $\gamma < 1$, so all sums converge absolutely. Rearranging this equality gives

$$D^{(\pi)} = (1-\gamma)D + \gamma P^{(\pi)}D^{(\pi)}. \tag{10}$$

It follows that

$$\sum_{a\in A}\Lambda_{s',a}^{(\pi)} = (1-\gamma)D_{s'} + \gamma\sum_{s\in S}\sum_{a\in A}\Lambda_{s,a}^{(\pi)}P_{s,a,s'}$$

for each $s' \in S$, where we have used the equalities

$$D_{s'}^{(\pi)} = \sum_{a\in A}\Lambda_{s',a}^{(\pi)}$$

and

$$\begin{aligned}(P^{(\pi)}D^{(\pi)})_{s'} &= \sum_{s\in S}P_{s,s'}^{(\pi)}D_s^{(\pi)}\\ &= \sum_{s\in S}\sum_{a\in A}P_{s,a,s'}\pi_{s,a}D_s^{(\pi)}\\ &= \sum_{s\in S}\sum_{a\in A}P_{s,a,s'}\Lambda_{s,a}^{(\pi)}\end{aligned}$$

that follow from the definition of $\Lambda^{(\pi)}$. Therefore, $\Lambda^{(\pi)}$ satisfies the first constraint.

Next, we show that $\Lambda^{(\pi)}$ satisfies the second constraint, which says that

$$\begin{aligned}\Bigg| & p_{\text{maj}}^{-1}\sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A}\lambda_{(\text{maj},\tilde{s}),a}\rho_{(\text{maj},\tilde{s}),a}\\ & - p_{\text{min}}^{-1}\sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A}\lambda_{(\text{min},\tilde{s}),a}\rho_{(\text{min},\tilde{s}),a}\Bigg| \le \epsilon.\end{aligned} \tag{11}$$

In particular, note that

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

since the value of z for s equals the value of z for the initial state $s_0 \sim D$. Furthermore, the probability of sampling $s \sim D^{(\pi)} \mid \exists\tilde{s}\in\tilde{S}\ .\ s = (z,\tilde{s})$ is

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

Together with the definition of $\Lambda_z^{(\pi)}$, we have

$$\begin{aligned}(\Lambda_z^{(\pi)})_{s,a} &= (D_z^{(\pi)})_s \pi_{s,a} \\ &= \frac{D_s^{(\pi)}\mathbb{I}[\exists \tilde{s} \in \tilde{S} \ . \ s = (z, \tilde{s})]}{p_z} \cdot \pi_{s,a} \\ &= \frac{\Lambda_{s,a}^{(\pi)}\mathbb{I}[\exists \tilde{s} \in \tilde{S} \ . \ s = (z, \tilde{s})]}{p_z}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}] \\ &= \sum_{s\in S}\sum_{a\in A} \frac{\Lambda_{s,a}^{(\pi)}\mathbb{I}[\exists \tilde{s} \in \tilde{S} \ . \ s = (z, \tilde{s})]}{p_z} \cdot \rho_{s,a} \\ &= p_z^{-1} \sum_{\tilde{s}\in\tilde{S}}\sum_{a\in A} \Lambda_{s,a}^{(\pi)}\rho_{s,a}. \qquad (12)\end{aligned}$$

By assumption, π satisfies the demographic parity constraint, which says exactly that (12) satisfies (11). Thus, $\Lambda^{(\pi)}$ satisfies the second constraint.

Therefore, $\Lambda^{(\pi)}$ is a feasible point of the LP, as claimed.

Step 2. Let $\lambda \in \mathbb{R}^{|S|\times|A|}$ be a feasible point of the LP in Algorithm 1, and let

$$\pi_{s,a} = \frac{\lambda_{s,a}}{\sum_{a'\in A}\lambda_{s,a'}}$$

be the corresponding policy returned by Algorithm 1. Then, we claim that $\lambda = \Lambda^{(\pi)}$, that $\pi \in \Pi_{\text{DP}}$, and that the value of the objective for λ equals $R^{(\pi)}$.

To see the first claim, let $d \in \mathbb{R}^{|S|}$ be defined by

$$d_s = \sum_{a\in A} \lambda_{s,a}.$$

We show that $D^{(\pi)} = d$. To this end, note that because λ satisfies the first constraint in the LP, we have

$$\sum_{a\in A} \lambda_{s',a} = (1-\gamma)D_{s'} + \gamma \sum_{s\in S}\sum_{a\in A} \lambda_{s,a}P_{s,a,s'}.$$

Together with the equality

$$\pi_{s,a} = \frac{\lambda_{s,a}}{d_s},$$

we have

$$\begin{aligned}d_{s'} &= (1-\gamma)D_{s'} + \gamma \sum_{s\in S}\sum_{a\in A} d_s\pi_{s,a}P_{s,a,s'} \\ &= (1-\gamma)D_{s'} + \gamma(P^{(\pi)}d)_{s'}.\end{aligned}$$

Thus,

$$d = (1-\gamma)D + \gamma P^{(\pi)}d. \qquad (13)$$

We note that $I - \gamma P^{(\pi)}$ is invertible—in particular, the eigenvalues of $\gamma P^{(\pi)}$ have norms bounded by γ, so the eigenvalues of $I - \gamma P^{(\pi)}$ have norms bounded below by $1 - \gamma$; therefore, the eigenvalues of $I - \gamma P^{(\pi)}$ are nonzero, so it is invertible. As a consequence, we can solve for d in (13) to get

$$d = (1-\gamma)(I - \gamma P^{(\pi)})^{-1}D.$$

Finally, from (10) in Step 1 of this proof, we established that $D^{(\pi)}$ similarly satisfies

$$D^{(\pi)} = (1-\gamma)D + \gamma P^{(\pi)}D^{(\pi)}.$$

As before, since $I - \gamma P^{(\pi)}$ is invertible, we have

$$D^{(\pi)} = (1-\gamma)(I - \gamma P^{(\pi)})^{-1}D = d.$$

Thus,

$$\lambda_{s,a} = d_s\pi_{s,a} = D_s\pi_{s,a} = \Lambda_{s,a}^{(\pi)},$$

so the first claim follows.

To see the second claim, note that since λ is feasible, it must satisfy the second constraint of the LP. As shown in the first step of this proof, (11) is equivalent to the demographic parity constraint. Thus, $\pi \in \Pi_{\text{DP}}$, as claimed.

To see the third claim, note that

$$\begin{aligned}R^{(\pi)} &= (1-\gamma)\mathbb{E}_{(s,a)\sim\Lambda^{(\pi)}}[R_{s,a}] \\ &= (1-\gamma)\sum_{s\in S}\sum_{a\in A} \Lambda_{s,a}^{(\pi)}R_{s,a}.\end{aligned}$$

In other words, the value of the objective of the LP for the point λ is equal to $R^{(\pi)}$, as claimed.

Step 3. Finally, we use the results from the previous two steps to prove the theorem statement. First, let π^* be the solution to (1). By the claim shown in the first step, $\Lambda^{(\pi^*)}$ is a feasible point of the LP in Algorithm 1. Furthermore, by the claim shown in the second step, the value of the objective for $\lambda = \Lambda^{(\pi^*)}$ is $R^{(\pi^*)}$.

Next, let λ^0 be the solution to the LP in Algorithm 1. By the claim shown in the second step, (i) $\lambda_0 = \Lambda^{(\pi_0)}$, where π_0 is the policy returned by Algorithm 1, (ii) $\pi_0 \in \Pi_{\text{DP}}$, and (iii) the value of the objective for λ^0 is $R^{(\pi_0)}$.

It follows that $R^{(\pi^*)} \leq R^{(\pi_0)}$, since π_0 maximizes the objective of the LP over feasible points (and $\Lambda^{(\pi^*)}$ is feasible). Since $\pi_0 \in \Pi_{\text{DP}}$, it follows that π_0 is also a solution to (1). Thus, we have proven the theorem statement. $\square$

D Proof of Theorem 4.2

Our proof proceeds in three steps. First, we bound the error $|\tilde{\rho}^{(\pi)} - \rho^{(\pi)}|$ due to truncation. Second, we bound the estimation error $|\hat{\rho}^{(\pi)} - \tilde{\rho}^{(\pi)}|$. Third, we combine steps 1 and 2 to prove Theorem 4.2.

Step 1. Note that for any policy π and any $z \in Z$, we have

$$\begin{aligned} |\tilde{\rho}_z^{(\pi)} - \rho_z^{(\pi)}| = \left| \sum_{t=T}^{\infty} \gamma^t \langle \rho_z, \Lambda^{(\pi,t)} \rangle \right| &\leq \sum_{t=T}^{\infty} \gamma^t \rho_{\max} \\ &\leq \frac{\gamma^T \rho_{\max}}{1-\gamma} \\ &\leq \frac{\sigma\epsilon}{4}. \end{aligned}$$

Step 2. For each $z \in Z$, let $\hat{\rho}_z^{(\pi)}$ be an estimate of $\tilde{\rho}_z^{(\pi)}$ using m sampled rollouts $\zeta^{(1)}, ..., \zeta^{(m)}$. First, note that

$$|\hat{\rho}_z^{(\pi)}| \leq \frac{\rho_{\max}}{1-\gamma}$$

is bounded, so we can apply Hoeffding's inequality (see Lemma F.1) to get

$$\begin{aligned} \Pr\left[|\hat{\rho}_z^{(\pi)} - \tilde{\rho}_z^{(\pi)}| \geq \frac{\sigma\epsilon}{4} \right] &\leq 2\exp\left(-\frac{m\sigma^2\epsilon^2}{32\rho_{\max}/(1-\gamma)} \right) \\ &\leq \frac{\delta}{2} \end{aligned}$$

Since $Z = \{\text{maj}, \text{min}\}$, by a union bound,

$$|\hat{\rho}_z^{(\pi)} - \tilde{\rho}_z^{(\pi)}| \leq \frac{\epsilon}{4} \qquad (\forall z \in Z)$$

with probability at least $1 - \delta$.

Step 3. Now, we can prove Theorem 4.2. First, note that with probability $1 - \delta$,

$$\begin{aligned} |\hat{\rho}_z^{(\pi)} - \rho_z^{(\pi)}| &\leq |\hat{\rho}^{(\pi)} - \tilde{\rho}^{(\pi)}| + |\tilde{\rho}^{(\pi)} - R^{(\pi)}| \\ &\leq \frac{\sigma\epsilon}{4} + \frac{\sigma^2\epsilon}{4} \\ &\leq \frac{\sigma\epsilon}{2}, \end{aligned}$$

for all $z \in Z$. Thus,

$$\begin{aligned} &|\rho_{\text{maj}}^{(\pi)} - \rho_{\text{min}}^{(\pi)}| \\ &\leq |\rho_{\text{maj}}^{(\pi)} - \hat{\rho}_{\text{maj}}^{(\pi)}| + |\hat{\rho}_{\text{maj}}^{(\pi)} - \hat{\rho}_{\text{min}}^{(\pi)}| + |\hat{\rho}_{\text{min}}^{(\pi)} - \rho_{\text{min}}^{(\pi)}| \\ &\leq \frac{\sigma\epsilon}{2} + (1-\sigma)\epsilon + \frac{\sigma\epsilon}{2} \\ &= \epsilon, \end{aligned}$$

which implies that $\pi \in \Pi_{\text{DP},\epsilon}$. Thus, the theorem follows. □

E Proof of Theorem 5.1

We prove the following lemma; Theorem 5.1 follows by choosing $\epsilon' = N^{-1/3}$.

Lemma E.1. *Let $\epsilon, \epsilon', \delta \in \mathbb{R}_+$ be given. Assume that R_{max} be an upper bound on R (i.e., $\|R\|_\infty = R_{max}$) and on ρ. Let $\tilde{\epsilon} = \min\{\epsilon, \epsilon'\}$, and let*

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

Let $\hat{M} = (S, A, D, \hat{P}, R, T)$, and $\hat{\pi}$ be the optimal policy for $\hat{M}$ in $\hat{\Pi}_{DP,\epsilon/2}$ (i.e., the set of policies satisfying demographic parity for $\hat{M}$). Let $M = (S, A, D, P, R, T)$, and π^* be optimal for M in $\Pi_{DP,\epsilon/4}$. Then, $\hat{\pi} \in \Pi_{DP,\epsilon}$, and $R^{(\pi^*)} - R^{(\hat{\pi})} \leq \epsilon'$, where $R^{(\pi)}$ is defined for M.

Our proof proceeds in three steps. First, we prove that for any ϵ_0, δ_0, we can choose N_0 sufficiently large so that

$$\|P - \hat{P}\|_\infty \leq \epsilon_0$$

with probability at least $1 - \delta_0$. Second, we prove that assuming $\|P - \hat{P}\|_\infty \leq \epsilon_0$, then for any policy π, we have

$$|R^{(\pi)} - \hat{R}^{(\pi)}| \leq T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0,$$

where $R^{(\pi)}$ (resp., $\hat{R}^{(\pi)}$) is the expected cumulative distribution assuming the transitions are P (resp., $\hat{P}$), and similarly for the agent rewards ρ. Third, we use the first two steps to prove the lemma statement.

Step 1. Given $\epsilon_0, \delta_0 \in \mathbb{R}_+$, we claim that for

$$N_0 = \frac{2\log(2|S|^2|A|/\delta_0)}{\lambda_0^2 \epsilon_0^2},$$

then our estimate $\hat{P}$ satisfies

$$\|\hat{P} - P\|_\infty \leq \epsilon_0$$

with probability at least $1 - \delta_0$.

Let $I_{s,a}$ be the random variable indicating whether our algorithm observes a tuple (s, a, s') (for some $s' \in S$) on a single episode, and let $I_{s,a,i}$ be samples of $I_{s,a}$ for each of the N_0 exploratory episodes taken by our algorithm. Let

$$\begin{aligned} \mu_{s,a}^{(I)} &= \mathbb{E}[I_{s,a}] \\ \hat{\mu}_{s,a}^{(I)} &= \frac{1}{N_0} \sum_{i=1}^{N_0} I_{s,a,i}. \end{aligned}$$

Then, by Hoeffding's inequality (see Lemma F.1), we have

$$\Pr\left[|\hat{\mu}_{s,a}^{(I)} - \mu_{s,a}^{(I)}| \geq \epsilon \right] \leq 2e^{-2N_0\epsilon^2}. \qquad (14)$$

By assumption, we have

$$\mu_{s,a}^{(I)} = \Lambda_{s,a}^{(\pi_0)} \geq \lambda_0,$$

so using $\epsilon = \lambda_0/2$ in (14), we have

$$\hat{\mu}_{s,a}^{(I)} \geq \frac{\mu_{s,a}^{(I)}}{2} \geq \frac{\lambda_0}{2} \tag{15}$$

with probability at least

$$1 - 2e^{-N_0(\mu_{s,a}^{(I)})^2/2} \geq 1 - 2e^{-N_0\lambda_0^2/2}.$$

Taking a union bound over $s \in S$ and $a \in A$, we have (15) holds for every $s \in S$ and $a \in A$ with probability at least

$$1 - 2|S| \cdot |A| \cdot e^{-N_0\lambda_0^2/2}. \tag{16}$$

In this event, we have at least $\frac{N_0\lambda_0}{2}$ observations (s, a, s') (for some $s' \in S$) for every $s \in S$ and $a \in A$.

Now, for an observation (s, a, s''), let $J_{s,a,s'}$ be the random variable indication whether $s' = s''$. Without loss of generality, we assume that we have exactly $N_1 = \frac{N_0\lambda_0}{2}$ samples $J_{s,a,s',j}$ of $J_{s,a,s'}$ for each $s \in S$ and $a \in A$. Let

$$\mu_{s,a,s'}^{(J)} = \mathbb{E}[J_{s,a,s'}]$$
$$\hat{\mu}_{s,a,s'}^{(J)} = \frac{1}{N_1}\sum_{j=1}^{N_1} J_{s,a,s',j}.$$

Then, by Hoeffding's inequality (see Lemma F.1), we have

$$\Pr\left[|\hat{\mu}_{s,a,s'}^{(J)} - \mu_{s,a,s'}| \geq \epsilon\right] \leq 2e^{-2N_1\epsilon^2}. \tag{17}$$

Note that by definition, $\mu_{s,a,s'}^{(J)} = P_{s,a,s'}$ and $\hat{\mu}_{s,a,s'}^{(J)} = \hat{P}_{s,a,s'}$. Thus, taking $\epsilon = \epsilon_0$ in (17), we have

$$|P_{s,a,s'} - \hat{P}_{s,a,s'}| \leq \epsilon_0 \tag{18}$$

with probability at least

$$1 - 2e^{-2N_1\epsilon_0^2}.$$

Taking a union bound over all $s, s' \in S$ and $a \in A$, we have (18) for all $s, s' \in S$ and $a \in A$ with probability at least

$$1 - 2|S|^2|A| \cdot e^{-2N_1\epsilon_0^2}. \tag{19}$$

In other words, in this event, we have $\|P - \hat{P}\|_\infty \leq \epsilon_0$.

Taking a union bound over (16) and (19), we have

$$\|P - \hat{P}\|_\infty \leq \epsilon_0$$

with probability at least

$$\begin{aligned}
&1 - 2|S|^2|A| \cdot e^{-2N_1\epsilon_0^2} - 2|S| \cdot |A| \cdot e^{-N_0\lambda_0^2/2} \\
&= 1 - 2|S|^2|A| \cdot e^{-N_0\lambda_0\epsilon_0^2} - 2|S| \cdot |A| \cdot e^{-N_0\lambda_0^2/2} \\
&\geq 1 - 2|S|^2|A| \cdot e^{-N_0\lambda_0^2\epsilon_0^2/2} \\
&= \delta_0,
\end{aligned}$$

as claimed.

Step 2. We claim that assuming

$$\|P - \hat{P}\|_\infty \leq \epsilon_0,$$

then for any policy π, we have

$$|R^{(\pi)} - \hat{R}^{(\pi)}| \leq T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0,$$

where $R^{(\pi)}$ is the expected cumulative reward for π in the MDP $M = (S, A, D, P, R, T)$ and $\hat{R}^{(\pi)}$ is the expected cumulative reward for π in the MDP $\hat{M} = (S, A, D, \hat{P}, R, T)$. Note that we have replaced the discount factor γ with the time horizon T. In addition, for all $z \in Z$, we have

$$|\rho_z^{(\pi)} - \hat{\rho}_z^{(\pi)}| \leq T \cdot |S| \cdot R_{\max} \cdot \epsilon_0,$$

where

$$\rho_z^{(\pi)} = \mathbb{E}_{(s,a)\sim\Lambda_z^{(\pi)}}[\rho_{s,a}],$$

is the expected cumulative agent reward for the MDP M, and $\hat{\rho}_z^{(\pi)}$ is the expected cumulative agent reward for the MDP $\hat{M}$. We only prove the claim for $|R^{(\pi)} - \hat{R}^{(\pi)}|$; the claim for $|\rho_z^{(\pi)} - \hat{\rho}_z^{(\pi)}|$ follows using the same argument.

Let $W \in \mathbb{R}^{|S|}$ be

$$W_s = \langle \pi_{s,\cdot}, R_{s,\cdot} \rangle = \sum_{a\in A} \pi_{s,a}R_{s,a}.$$

Then, we have

$$\begin{aligned}
R^{(\pi)} &= \langle R, \Lambda^{(\pi)} \rangle \\
&= \sum_{s\in S}\sum_{a\in A} D_s^{(\pi)}\pi_{s,a}R_{s,a} \\
&= \langle D^{(\pi)}, W \rangle.
\end{aligned}$$

Now, note that

$$D^{(\pi,t)} = (P^{(\pi)})D,$$

so we have

$$\begin{aligned}
D^{(\pi)} &= \frac{1}{T}\sum_{t=0}^{T-1} D^{(\pi,t)} \\
&= \frac{1}{T}\left[\sum_{t=0}^{T-1}(P^{(\pi)})^t D\right].
\end{aligned}$$

Thus,

$$R^{(\pi)} = \sum_{t=0}^{T-1} \left\langle (P^{(\pi)})^t D, W \right\rangle .$$

Similarly,

$$\hat{R}^{(\pi)} = \sum_{t=0}^{T-1} \left\langle (\hat{P}^{(\pi)})^t D, W \right\rangle .$$

It follows that

$$R^{(\pi)} - \hat{R}^{(\pi)} = \sum_{t=0}^{T-1} \left\langle (P^{(\pi)})^t D - (\hat{P}^{(\pi)})^t D, W \right\rangle .$$

Thus,

$$\begin{aligned} & |R^{(\pi)} - \hat{R}^{(\pi)}| \\ & \leq \sum_{t=0}^{T-1} \|(P^{(\pi)})^t D - (\hat{P}^{(\pi)})^t D\|_\infty \cdot \|W\|_1 \\ & \leq \sum_{t=1}^{T-1} \epsilon_0 \cdot T \cdot \|D\|_\infty \cdot \|W\|_1, \end{aligned} \tag{20}$$

where the first line follows from Hölder's inequality, and the second line follows from properties of the matrix norm, from the fact that

$$\begin{aligned} & \|(P^{(\pi)})^t - (\hat{P}^{(\pi)})^t\|_\infty \\ & = \|P^{(\pi)} - \hat{P}^{(\pi)}\|_\infty \cdot \sum_{s=0}^{t-1} \|P^{(\pi)}\|_\infty^s \cdot \|\hat{P}^{(\pi)}\|_\infty^{t-s-1} \\ & \leq \epsilon_0 \cdot T, \end{aligned}$$

and using the fact that the summand is zero for $t = 0$ since $(P^{(\pi)})^0 = (\hat{P}^{(\pi)})^0 = I$. Note that

$$\|D\|_\infty \leq 1 \tag{21}$$

Furthermore,

$$\begin{aligned} |W_s| = |\langle \pi_{s,\cdot}, R_{s,\cdot} \rangle| & \leq \|\pi_{s,\cdot}\|_1 \cdot \|R_{s,\cdot}\|_\infty \\ & \leq \|R_{s,\cdot}\|_\infty \\ & \leq R_{\max}, \end{aligned}$$

where the first inequality follows from Hölder's inequality and the second inequality follows since $\pi_{s,\cdot}$ is a discrete probability distribution. Therefore,

$$\|W\|_1 = \sum_{s \in S} |W_s| \leq |S| \cdot R_{\max}. \tag{22}$$

Plugging (21) and (22) into (20) gives

$$|R^{(\pi)} - \hat{R}^{(\pi)}| \leq T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0.$$

Step 3. Now, we prove the theorem. Let $\hat{\pi}$ be the optimal policy for $\hat{M}$ (i.e., transitions $\hat{P}$) satisfying $\hat{\pi} \in \hat{\Pi}_{\mathrm{DP},\epsilon/2}$. Similarly, let π^* be the optimal policy for M (i.e., transitions P) satisfying $\pi^* \in \Pi_{\mathrm{DP},\epsilon/4}$. We apply the second step with

$$\begin{aligned} \epsilon_0 &= \frac{\tilde{\epsilon}}{8T^2 \cdot |S| \cdot R_{\max}} \\ \delta_0 &= \delta. \end{aligned}$$

Then, by the first step, for all $z, z' \in Z$, we have

$$\begin{aligned} & \rho_z^{(\hat{\pi})} - \rho_{z'}^{(\hat{\pi})} \\ & \leq (\rho_z^{(\hat{\pi})} - \hat{\rho}_z^{(\hat{\pi})}) + (\hat{\rho}_z^{(\hat{\pi})} - \hat{\rho}_{z'}^{(\hat{\pi})}) + +(\rho_{z'}^{(\hat{\pi})} - \hat{\rho}_{z'}^{(\hat{\pi})}) \\ & \leq T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0 + \frac{\epsilon}{2} + T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0 \\ & \leq \epsilon, \end{aligned}$$

where the inequality on the third line follows because $\hat{\pi} \in \hat{\Pi}_{\mathrm{DP},\epsilon/2}$, and the inequality on the last line follows since $\tilde{\epsilon} \leq \epsilon$. Thus, we guarantee that $\hat{\pi} \in \Pi_{\mathrm{DP},\epsilon}$.

Next, note that similarly, for all $z, z' \in Z$, we have

$$\hat{\rho}_z^{(\pi^*)} - \hat{\rho}_{z'}^{(\pi^*)} \leq \frac{\epsilon}{2},$$

so $\pi^* \in \hat{\Pi}_{\mathrm{DP},\epsilon/2}$. As a consequence, we have

$$\begin{aligned} & R^{(\pi^*)} - R^{(\hat{\pi})} \\ & = (R^{(\pi^*)} - \hat{R}^{(\pi^*)}) + (\hat{R}^{(\pi^*)} - \hat{R}^{(\hat{\pi})}) + (\hat{R}^{(\hat{\pi})} - R^{(\hat{\pi})}) \\ & \leq T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0 + 0 + T^2 \cdot |S| \cdot R_{\max} \cdot \epsilon_0 \\ & \leq \epsilon', \end{aligned}$$

where the inequality on the third line follows because $\hat{\pi}$ maximizes $\hat{R}^{(\pi)}$ over $\pi \in \hat{\Pi}_{\mathrm{DP},\epsilon/2}$ and $\pi^* \in \hat{\Pi}_{\mathrm{DP},\epsilon/2}$, and the inequality on the last line follows since $\tilde{\epsilon} \leq \epsilon'$.

Thus, the lemma statement follows. □

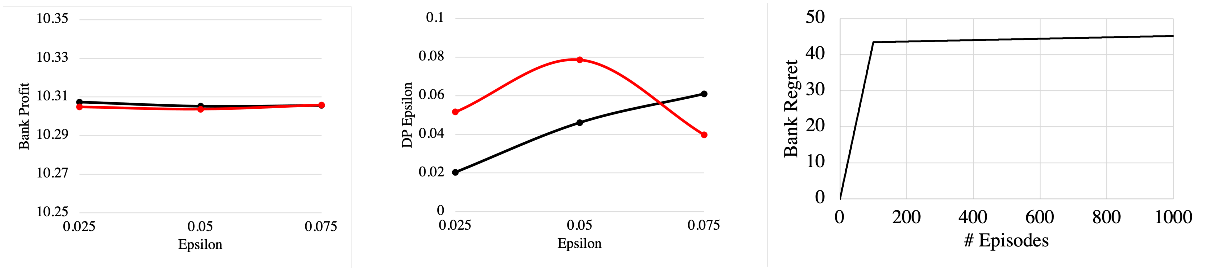
F Technical Lemmas

Lemma F.1. *(Hoeffding's inequality) Let $X \sim p_X$ be a random variable with domain $[a, b] \subseteq \mathbb{R}$ and mean μ_X, and let $\hat{\mu}_X = n^{-1} \sum_{i=1}^n X_i$ be an estimate of μ_X a using n i.i.d. samples $X_i \sim p_X$. Then, we have*

$$Pr[|\hat{\mu}_X - \mu_X| \geq \epsilon] \leq 2 \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right), \tag{23}$$

where the probability is taken over the randomness in the i.i.d. samples $X_1, ..., X_n \sim p_X$.

Proof. See Wainwright (2019) for a proof. □



(a) (b) (c)

Figure 2: Demographic parity (a) objective value, (b) constraint value for our algorithm (black) and the optimistic baseline (red). (c) Regret of our reinforcement learning algorithm.

G Experimental Details & Additional Results

Parameters. We use the following parameters for our loan MDP:

$$\begin{aligned}
I &= 0.17318629 \\
p_Z &= 0.29294318 \\
\alpha_{\text{maj}} &= 0.65338681 \\
\beta_{\text{maj}} &= 0.20783559 \\
\alpha_{\text{min}} &= 0.48824268 \\
\beta_{\text{min}} &= 0.48346869 \\
\lambda &= 0.01 \\
\tau &= 0.1 \\
\epsilon &= 0.1 \\
T &= 50 \\
T_{\text{maj}} &= 10 \\
T_{\text{min}} &= 7.
\end{aligned}$$

Additional results for Algorithm 2. We additionally study how Algorithm 2 varies with the fairness constraint threshold ϵ. In Figure 2 (a,b), we show the objective value achieved and the fairness constraint value achieved by our algorithm and the optimistic algorithm for the demographic parity constraint. While the objective values achieved are very similar, the optimistic algorithm does not always satisfy the fairness constraint. In particular, for $\epsilon = 0.025$, its constraint value is 0.052 (exceeds ϵ by 108%), and for $\epsilon = 0.05$, it is 0.079 (exceeds ϵ by 59%). Intuitively, there are multiple policies that achieve the same objective value, but the optimistic algorithm sometimes fails to find the ones that are fair. In contrast, our algorithm always satisfies the fairness constraint.

Results for Algorithm 1. We have evaluated Algorithm 1 on a modified version of our loan MDP where α and β are discretized and thresholded to make the state space finite. For this MDP, we have compared Algorithm 1 to solving an unconstrained MDP—i.e., without the demographic parity fairness constraint. We use $\epsilon = 0.01$. Our results are as follows: (i) for Algorithm 1, the cumulative expected reward is 0.68 and the fairness constraint value is 0.01, and (ii) for the unconstrained algorithm, the cumulative expected reward is 0.69 and the fairness constraint value is 0.26. In other words, for a small reduction in reward, our algorithm substantially improves fairness. The remaining baselines cannot be implemented using the approach in Algorithm 1.

Results for reinforcement learning. We have run our reinforcement learning algorithm in conjunction with the We run the algorithm for $N = 1000$ episodes total. In particular, we explore for 100 episodes using a conservative policy π_0 that ignores the state; then, we use the estimated transitions to learn the optimal policy $\hat{\pi}$ and use $\hat{\pi}$ for the remaining 900 episodes. Since our model is parameterized by α and β, we estimate these quantities instead of directly estimating the transitions. We show the regret compared to the optimal policy in Figure 2 (c), averaged over 5 iterations. As can be seen, the regret quickly increases while using π_0, and then becomes almost flat when using $\hat{\pi}$. We note that our algorithm satisfies the fairness constraint across all episodes and iterations.