
On Welfare-Centric Fair Reinforcement Learning

Supplementary Appendices

A Proof Compendium

We now compute the per-state sample complexity for an (α, β) approximation of the MDP \mathcal{M} .

Lemma 4.3 (Per-State Sample Complexity). *Suppose MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{R}, \mathbf{P}, \gamma \rangle$, and let*

$$m_{\text{knw}} \doteq \left\lceil \ln \left(\frac{|\mathcal{S}||\mathcal{A}|(2^{|\mathcal{S}|} - 2 + 2g)}{\delta} \right) \max \left(\frac{1}{2\alpha^2}, \frac{R_{\max}}{2\beta^2} \right) \right\rceil, \quad (4)$$

where $R_{\max} \doteq \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\mathbf{R}(s, a)\|_{\infty} \in [0, \infty]$. Now if $\hat{\mathcal{M}}$ is estimated from m_{knw} samples of each state-action pair, then, with probability at least $1 - \delta$, $\hat{\mathcal{M}}$ is an α - β approximation of \mathcal{M} .

Proof. This result is an adaptation of proposition 2.1 of (Agarwal et al., 2022). For any state s and action a , let $\mathbf{E}_{s,a,m,S}$ and $\mathbf{E}_{s,a,m,R}$ represent the corresponding next-state and reward buffers of length m . Let $\mathbf{e}_s \in \{0, 1\}^{|\mathcal{S}|}$ such that $(\mathbf{e}_s)_s = 1$ and all other entries of \mathbf{e}_s are zero.

$$\begin{aligned} \hat{\mathbf{P}}(s, a) &= \frac{1}{m} \sum_{s' \in \mathbf{E}_{s,a,m,S}} \mathbf{e}_{s'} \\ \forall i \in 1, \dots, g : \hat{\mathbf{R}}_i(s, a) &= \frac{1}{m} \sum_{\mathbf{r} \in \mathbf{E}_{s,a,m_{s,a},R}} r_i. \end{aligned} \quad (5)$$

$\hat{\mathbf{P}}$ and $\hat{\mathbf{R}}$ are maximum likelihood estimates of the transition and reward functions. Furthermore, for any vector $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$, we can write

$$\|\mathbf{v}\|_1 = \sup_{\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}} \mathbf{u} \cdot \mathbf{v}. \quad (6)$$

Thus, we can bound the $\|\cdot\|_1$ error in the transition probabilities as

$$\begin{aligned} \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \|\hat{\mathbf{P}}(s, a) - \hat{\mathbf{P}}(s, a)\|_1 &= \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \max_{\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}} \mathbf{u} \cdot (\hat{\mathbf{P}}(s, a) - \mathbf{P}(s, a)) \\ &= \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \max_{\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}} \mathbf{u} \cdot \frac{1}{m} \sum_{j=1}^m (\mathbf{e}_{s'}^j - \mathbf{P}(s, a)) \\ &= \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \max_{\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}} \frac{1}{m} \sum_{j=1}^m \mathbf{u} \cdot (\mathbf{e}_{s'}^j - \mathbf{P}(s, a)) \\ &= \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \max_{\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}} \frac{1}{m} \sum_{j=1}^m \mathbf{u} \cdot \mathbf{e}_{s'}^j - \frac{1}{m} \sum_{j=1}^m \mathbf{u} \cdot \mathbf{P}(s, a). \end{aligned} \quad (7)$$

Here $\mathbf{u} \cdot \mathbf{e}_{s'}$ is a random variable with range 1. Applying Hoeffding's bound at confidence level $\delta \in (0, 1)$ and taking a union over all (s, a) pairs as well as overall $\mathbf{u} \in \{-1, 1\}^{|\mathcal{S}|}$, we get with $1 - \delta$ probability,

$$\max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \|\hat{\mathbf{P}}(s, a) - \hat{\mathbf{P}}(s, a)\|_1 \leq \sqrt{\frac{1}{2m} \ln \left(\frac{|\mathcal{S}||\mathcal{A}|(2^{|\mathcal{S}|} - 2)}{\delta} \right)}. \quad (8)$$

Note that we have excluded $\mathbf{u} = -\mathbf{1}$ and $\mathbf{u} = \mathbf{1}$ as they do not contribute to the bound.

Similarly, we can bound the $\|\cdot\|_\infty$ error in the empirical reward vectors as

$$\max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \|\hat{\mathbf{R}}(s, a) - \hat{\mathbf{R}}(s, a)\|_\infty = \max_{i \in 1, \dots, g} \max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} |\hat{\mathbf{R}}_i(s, a) - \mathbf{R}_i(s, a)|. \quad (9)$$

Applying the two-tail Hoeffding's bound and taking a union over all (s, a, i) tuples, we get with $1 - \delta$ probability,

$$\max_{\substack{s \in \mathcal{S}, \\ a \in \mathcal{A}}} \|\hat{\mathbf{R}}(s, a) - \hat{\mathbf{R}}(s, a)\|_\infty \leq R_{\max} \sqrt{\frac{1}{2m} \ln \left(\frac{|\mathcal{S}||\mathcal{A}|2g}{\delta} \right)}. \quad (10)$$

Combining results in (7) and (10), and using the fact that $\forall (s, a, i)$ tuples, $\text{TVD}(T'(\cdot|s, a), T(\cdot|s, a)) \leq \alpha$, $|\mathbf{R}'_i(s, a) - \mathbf{R}_i(s, a)| \leq \beta$, we get with $1 - \delta$ probability,

$$m_{\text{knw}} \doteq \left\lceil \ln \left(\frac{|\mathcal{S}||\mathcal{A}|(2^{|\mathcal{S}|} - 2 + 2g)}{\delta} \right) \max \left(\frac{1}{2\alpha^2}, \frac{R_{\max}}{2\beta^2} \right) \right\rceil. \quad (11)$$

□

Lemma A.1 (Simulation). *Suppose that $\mathcal{M}' = \langle S, A, \mathbf{P}', \mathbf{R}', \gamma \rangle$ is an (α, β) uniform approximation of a vector-reward MDP $\mathcal{M} = \langle S, A, \mathbf{P}, \mathbf{R}, \gamma \rangle$. For any policy $\pi \in \Pi$ and group $i \in 1, \dots, g$, let \mathbf{V}_i^π and $\mathbf{V}_i'^\pi$ denote the value function of the policy π in \mathcal{M} and \mathcal{M}' respectively. Then,*

$$\|\mathbf{V}_i^\pi - \mathbf{V}_i'^\pi\|_\infty \leq \frac{2\beta + \gamma\alpha R_{\max}}{2(1 - \gamma)^2}. \quad (12)$$

Furthermore, for any $\varepsilon \in \mathbb{R}_+$, setting $\beta = \alpha\sqrt{R_{\max}}$ and $\alpha = 2\varepsilon(1 - \gamma)^2 / (2\sqrt{R_{\max}} + \gamma R_{\max})$ yields

$$\|\mathbf{V}_i^\pi - \mathbf{V}_i'^\pi\|_\infty \leq \varepsilon. \quad (13)$$

Proof. This result is due to Agarwal et al. (2022), and is an improvement over the classical Simulation Lemma (Lemma 1 in (Strehl et al., 2009)).

For any policy π and group $i \in 1, \dots, g$, we will use shorthand V^π to represent \mathbf{V}_i^π . We will prove the result for a single group and the same result directly follows for the rest of the groups.

For any state $s \in \mathcal{S}$ and policy $\pi \in \Pi$,

$$\begin{aligned} \|V'^\pi - V^\pi\|_\infty &= \max_{s \in \mathcal{S}} |V'^\pi(s) - V^\pi(s)| \\ &= \left| \sum_{a \in \mathcal{A}} \pi(s, a) \left(R'(s, a) + \gamma \mathbf{P}'(s, a)^\top V'^\pi - R(s, a) - \gamma \mathbf{P}(s, a)^\top V^\pi \right) \right| \\ &\leq \beta + \gamma \left| \sum_{a \in \mathcal{A}} \pi(s, a) \left(\mathbf{P}'(s, a)^\top V'^\pi - \mathbf{P}(s, a)^\top V'^\pi + \mathbf{P}(s, a)^\top V'^\pi - \mathbf{P}(s, a)^\top V^\pi \right) \right| \\ &\leq \beta + \gamma \max_{a \in \mathcal{A}} |(\mathbf{P}(s, a) - \mathbf{P}'(s, a))^\top V'^\pi| + \gamma \|V^\pi - V'^\pi\|_\infty \\ &\stackrel{(a)}{=} \beta + \gamma \max_{a \in \mathcal{A}} \left| (\mathbf{P}(s, a) - \mathbf{P}'(s, a))^\top \left(V'^\pi - \frac{R_{\max}}{(1 - \gamma)2} \cdot \mathbf{1} \right) \right| + \gamma \|V^\pi - V'^\pi\|_\infty \\ &\leq \beta + \gamma \max_{a \in \mathcal{A}} \|(\mathbf{P}(s, a) - \mathbf{P}'(s, a))\|_\infty \left\| V'^\pi - \frac{R_{\max}}{(1 - \gamma)2} \cdot \mathbf{1} \right\|_1 + \gamma \|V^\pi - V'^\pi\|_\infty \\ &\leq \beta + \frac{\gamma\alpha R_{\max}}{2(1 - \gamma)} + \gamma \|V'^\pi - V^\pi\|_\infty \\ \|V'^\pi - V^\pi\|_\infty &\leq \frac{\beta}{1 - \gamma} + \frac{\gamma\alpha R_{\max}}{2(1 - \gamma)^2} \\ &\leq \frac{2\beta + \gamma\alpha R_{\max}}{2(1 - \gamma)^2}. \end{aligned} \quad (14)$$

Note that we subtract $\frac{V_{\max}}{2} \cdot \mathbf{1} = \frac{R_{\max}}{2(1-\gamma)} \cdot \mathbf{1}$ from the value function V'^π to center its range around the origin, which uses the fact that both $\mathbf{P}(s, a)$ and $\mathbf{P}'(s, a)$ are transition probabilities that sum to 1.

The second result follows from simple algebraic manipulations. \square

Lemma A.2 (Simulation under T -E Reachability). *Suppose MDPs $\mathcal{M}, \hat{\mathcal{M}}$ with transition functions $\mathbf{P}, \hat{\mathbf{P}}$ and reward functions $\mathbf{R}, \hat{\mathbf{R}}$, and some known set $\mathcal{S}' \subseteq \mathcal{S}$ such that for all $s \in \mathcal{S}'$ and $a \in \mathcal{A}$, it holds that $\text{TVD}(\mathbf{P}'(\cdot|s, a), \mathbf{P}(\cdot|s, a)) \leq \alpha$, and $\|\mathbf{R}'(s, a) - \mathbf{R}(s, a)\|_\infty \leq \beta$, i.e., $\hat{\mathcal{M}}$ (α, β) -approximates \mathcal{M} over \mathcal{S}' . Suppose also some state s such that no policy in $\hat{\mathcal{M}}$ can reach $\mathcal{S} \setminus \mathcal{S}'$ within T steps with probability at least E . For any policy π , let \mathbf{V}^π and $\hat{\mathbf{V}}^\pi$ be the value functions of π in \mathcal{M} and $\hat{\mathcal{M}}$, respectively. Then*

$$\|\hat{\mathbf{V}}^\pi(s) - \mathbf{V}^\pi(s)\|_\infty \leq \frac{1}{1-\gamma} \left(\frac{2\beta + \alpha\gamma R_{\max}}{2(1-\gamma)} + (E + \alpha T)R_{\max} + \gamma^T R_{\max} \right). \quad (15)$$

Proof. Without loss of generality, we will prove the upper bound on the error in the value functions of one group, i.e., we will assume scalar rewards. The claim in the lemma will then directly follow from this result.

Let p represent the probability of reaching $\mathcal{S} \setminus \mathcal{S}'$ from state s while following policy π for T steps in \mathcal{M} , i.e., $p = \mathbb{P}(\bigvee_{i=1}^T s_i \in \mathcal{S} \setminus \mathcal{S}' \mid \pi, s_0 = s, s_t = \mathbf{P}_{s_{t-1}})$.

$$\begin{aligned} |\hat{\mathbf{V}}^\pi(s) - \mathbf{V}^\pi(s)| &= |\hat{\mathbf{V}}^\pi(s) - \tilde{\mathbf{V}}^\pi(s) + \tilde{\mathbf{V}}^\pi(s) - \mathbf{V}^\pi(s)| \\ &\leq |\hat{\mathbf{V}}^\pi(s) - \tilde{\mathbf{V}}^\pi(s)| + |\tilde{\mathbf{V}}^\pi(s) - \mathbf{V}^\pi(s)| && \text{Triangle Inequality} \\ &\leq \frac{2\beta + \alpha\gamma R_{\max}}{2(1-\gamma)^2} + |\tilde{\mathbf{V}}^\pi(s) - \mathbf{V}^\pi(s)| && \text{Lemma A.1} \\ &\leq \frac{2\beta + \alpha\gamma R_{\max}}{2(1-\gamma)^2} + \frac{pR_{\max}}{1-\gamma} + (1-p)\frac{\gamma^T R_{\max}}{1-\gamma} && \text{See Below} \\ &\leq \frac{1}{1-\gamma} \left(\frac{2\beta + \alpha\gamma R_{\max}}{2(1-\gamma)} + (E + \alpha T)R_{\max} + \gamma^T R_{\max} \right) && \begin{array}{l} p \leq E + \alpha T \\ 1-p \leq 1 \end{array} \end{aligned}$$

The step marked *See Below* follows from the fact that MDP $\hat{\mathcal{M}}(\alpha, \beta)$ approximates MDP \mathcal{M} and therefore, with probability at most $p \leq E + \alpha T$, policy π will escape \mathcal{S}' in \mathcal{M} before the completion of T steps and gain at most $\frac{R_{\max}}{1-\gamma}$ additional returns, and with probability at least $(1-p) \leq 1 - (E + \alpha T)$ policy π will escape \mathcal{S}' in \mathcal{M} after T steps and gain at most $\frac{\gamma^T R_{\max}}{1-\gamma}$ additional returns. \square

Lemma A.3 (Welfare Approximation Error). *Suppose that for any state $s \in \mathcal{S}$,*

$$\sup_{\pi \in \Pi_{\mathcal{M}}} \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty \leq \frac{\varepsilon}{2\lambda}.$$

Then, any $\lambda\|\cdot\|_\infty$ Lipschitz welfare function $W(\cdot)$ satisfies

$$\forall \pi : \left| W(\mathbf{V}^\pi(s)) - W(\hat{\mathbf{V}}^\pi(s)) \right| \leq \frac{\varepsilon}{2}.$$

Hence, the empirical-welfare-optimal policy $\hat{\pi} \doteq \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} W(\hat{\mathbf{V}}^\pi(s))$ obeys

$$W(\mathbf{V}^{\hat{\pi}}(s)) \geq \sup_{\pi^* \in \Pi_{\mathcal{M}}} W(\mathbf{V}^{\pi^*}(s)) - \varepsilon.$$

Proof. The first result simply follows from the Lipschitz property of the welfare function $W(\cdot)$. Since W is $\lambda\|\cdot\|_\infty$ Lipschitz, we can write

$$\forall \pi : |W(\mathbf{V}^\pi(s)) - W(\hat{\mathbf{V}}^\pi(s))| \leq \lambda \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty. \quad (16)$$

Then, based on the assumption $\sup_{\pi \in \Pi_{\mathcal{M}}} \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty \leq \frac{\varepsilon}{2\lambda}$, we obtain

$$\forall \pi : \left| W(\mathbf{V}^\pi(s)) - W(\hat{\mathbf{V}}^\pi(s)) \right| \leq \frac{\varepsilon}{2} . \quad (17)$$

Suppose that π^* is welfare-optimal, i.e., $\pi^* \doteq \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} W(\mathbf{V}^\pi(s))$. Then,

$$\begin{aligned} & |W(\mathbf{V}^{\pi^*}(s)) - W(\mathbf{V}^{\hat{\pi}}(s))| \\ &= |W(\mathbf{V}^{\pi^*}(s)) - W(\hat{\mathbf{V}}^{\pi^*}(s)) + W(\hat{\mathbf{V}}^{\pi^*}(s)) - W(\mathbf{V}^{\hat{\pi}}(s))| \\ &\leq |W(\mathbf{V}^{\pi^*}(s)) - W(\hat{\mathbf{V}}^{\pi^*}(s)) + W(\hat{\mathbf{V}}^{\hat{\pi}}(s)) - W(\mathbf{V}^{\hat{\pi}}(s))| && \hat{\pi} \text{ is welfare-optimal in } \hat{\mathcal{M}} \\ &\leq |W(\mathbf{V}^{\pi^*}(s)) - W(\hat{\mathbf{V}}^{\pi^*}(s))| + |W(\hat{\mathbf{V}}^{\hat{\pi}}(s)) - W(\mathbf{V}^{\hat{\pi}}(s))| && \text{Triangle Inequality} \\ &\leq \lambda \|\mathbf{V}^{\pi^*}(s) - \hat{\mathbf{V}}^{\pi^*}(s)\|_\infty + \lambda \|\hat{\mathbf{V}}^{\hat{\pi}}(s) - \mathbf{V}^{\hat{\pi}}(s)\|_\infty && \lambda \cdot \|\cdot\|_\infty \text{ property of } W(\cdot) \\ &\leq \lambda \cdot \frac{\varepsilon}{2\lambda} + \lambda \cdot \frac{\varepsilon}{2\lambda} && \sup_{\pi \in \Pi_{\mathcal{M}}} \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty \leq \frac{\varepsilon}{2\lambda} \\ &\leq \varepsilon . \end{aligned}$$

□

Lemma A.4 (Explore or Exploit). *Suppose the empirical MDP $\hat{\mathcal{M}}$ estimated by algorithm 1 at any timestep (α, β) -approximates \mathcal{M} over $\mathcal{S} \setminus \mathcal{S}_{\text{unk}}$, and $\frac{6\lambda R_{\max}}{\varepsilon(1-\gamma)} > 1$ and $E \in (0, 1)$ holds. For any state $s \in \mathcal{S}$, let escape policy π_{esc} be the T -step temporal policy that maximizes the probability of reaching \mathcal{S}_{unk} from s , i.e., $\pi_{\text{esc}} \doteq \operatorname{argmax}_{\pi \in \Pi_T} \mathbb{P}(s_T \in \mathcal{S}_{\text{unk}} \mid s_0 = s, s_t \sim \hat{\mathbf{P}}(s_{t-1}, \pi(s_{t-1}, t)))$ and let p_{esc} be escape probability of π_{esc} . Let welfare-optimal policy $\pi^* \doteq \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} W(\mathbf{V}^\pi(s))$ and exploit policy $\pi_{\text{xplt}} \doteq \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} W(\hat{\mathbf{V}}^\pi(s))$. If $p_{\text{esc}} \geq E$, then, executing π_{esc} reaches \mathcal{S}_{unk} within T steps with probability at least $\frac{E}{2}$, otherwise π_{xplt} is ε -welfare-optimal, i.e., $W(\mathbf{V}^{\pi_{\text{xplt}}}(s)) \geq W(\mathbf{V}^{\pi^*}(s)) - \varepsilon$.*

Proof. We will prove this result by deriving the values of α , β , and T , and note that these are the values that are used in algorithm 1.

First, we will show that for threshold E , when $p_{\text{esc}} < E$, then the optimal policy is ε -welfare-optimal. We lay out the boundary conditions that are necessary but not sufficient to prove this result.

$$\begin{aligned} (1) & E > \alpha T \\ (2) & T \in \mathbb{Z}_+ \\ (3) & 0 < E < 1 . \end{aligned} \quad (18)$$

The first condition is necessary to ensure that the escape probability is positive and that the lower-bound on the escape probability, i.e., $E - \alpha T$ is also positive, the second condition is necessary to allow escape to \mathcal{S}_{unk} when there does not exist any ε -welfare-optimal exploit policy, and the third condition ensures that the escape probability threshold E is valid, i.e., it lies in the range $(0, 1)$. If $E = 1$ then the agent always explores, and if $E = 0$, then, the agent always exploits, thereby making it infeasible for the agent to learn and guarantee that all exploit policies are ε -welfare-optimal.

From lemma 4.3, we know that for any state $s \in \mathcal{S}$, if $\sup_{\pi \in \Pi_{\mathcal{M}}} \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty \leq \frac{\varepsilon}{2\lambda}$, then, π_{xplt} is ε -welfare-optimal, i.e.,

$$W(\mathbf{V}_1^{\pi^*}(s), \dots, \mathbf{V}_g^{\pi^*}(s)) - W(\mathbf{V}_1^{\pi_{\text{xplt}}}(s), \dots, \mathbf{V}_g^{\pi_{\text{xplt}}}(s)) \leq \varepsilon . \quad (19)$$

Note that $\hat{\pi} = \pi_{\text{xplt}}$ in lemma A.3.

From lemma A.2, we can write,

$$\sup_{\pi \in \Pi_{\mathcal{M}}} \|\mathbf{V}^\pi(s) - \hat{\mathbf{V}}^\pi(s)\|_\infty \leq \frac{1}{1-\gamma} \left(\frac{2\beta + \alpha\gamma R_{\max}}{2(1-\gamma)} + (E + \alpha T)R_{\max} + \gamma^T R_{\max} \right) . \quad (20)$$

Note that in algorithm 1, the escape probability of any exploit policy π is upper-bounded by $E + \alpha T$ in \mathcal{M} .

Therefore, combining (19) and (20), we obtain that π_{xplt} is ε -welfare-optimal if

$$\frac{R_{\max}}{1-\gamma} \left(\frac{\frac{2\beta}{R_{\max}} + \alpha\gamma}{2(1-\gamma)} + (E + \alpha T) + \gamma^T \right) \leq \frac{\varepsilon}{2\lambda} . \quad (21)$$

We need to find values of T , E , α , β such that they satisfy the equations in (21) and (18).

First, we set $E = 2\alpha T$ so that the boundary condition $E > \alpha T$ is satisfied and the lower bound on the escape probability $E - \alpha T$ is positive. Next, we set $\beta = \alpha\sqrt{R_{\max}}$. This is the optimal value of β in lemma 4.3. Note that these assignments for E and β do not violate the boundary conditions in (18). Then, we find the smallest T that satisfies $\gamma^T \leq \frac{\varepsilon(1-\gamma)}{6\lambda R_{\max}}$.

This gives us

$$T = \left\lceil \log_{\frac{1}{\gamma}} \left(\frac{6\lambda R_{\max}}{\varepsilon(1-\gamma)} \right) \right\rceil , \quad (22)$$

with an additional boundary condition that $\frac{6\lambda R_{\max}}{\varepsilon(1-\gamma)} > 1$. Note that we assume this condition holds in the lemma statement.

Substituting the values of T , β and E in (21), we get

$$\alpha \frac{R_{\max}}{1-\gamma} \left(\frac{\frac{2}{\sqrt{R_{\max}}} + \gamma}{2(1-\gamma)} + 3T \right) \leq \frac{\varepsilon}{3\lambda} . \quad (23)$$

Solving for α , we get

$$\alpha = \frac{\frac{\varepsilon(1-\gamma)}{3\lambda R_{\max}}}{\left(\frac{\frac{2}{\sqrt{R_{\max}}} + \gamma}{2(1-\gamma)} \right) + 3T} . \quad (24)$$

Finally, we obtain E by substituting the values of α and T in $E = 2\alpha T$.

In the case where $p_{\text{esc}} \geq E$, π_{esc} will result in reaching \mathcal{S}_{unk} in \mathcal{M} with probability atleast $E - \alpha T = E - \frac{E}{2} = \frac{E}{2}$. This follows from the fact that the transition probabilities of $\hat{\mathcal{M}}$ are α -TVD approximation of the true transition probabilities \mathbf{P} and therefore, we may incur at most αT error in the T-step escape probability of the exploit policy. \square

A.1 Handling failure probabilities of \mathbf{E}^4

Our main result, theorem 4.4, guarantees that algorithm 1 achieves an ε -optimal-welfare performance with probability at least $1 - \delta$.

There are two sources of failure for this algorithm. The first one is the error in the approximation of the rewards and next-state transition probabilities for a known state $s \in \mathcal{S} \setminus \mathcal{S}_{\text{unk}}$. The second is the set of attempted explorations that may fail to generate sufficient balanced wandering steps to result in a new known state.

Similar to (Kearns and Singh, 2002), we handle these failures by allowing each of these failures to occur with at most $\delta/2$ probability.

The first probability of the failure can be controlled using Lemma 4.3. To address the second source of failure, we treat each exploration attempt as a Bernoulli random variable with a probability of success being at least $E - \alpha T = \frac{E}{2}$, as derived in Lemma A.4. This probability of success indicates that the exploration policy π_{esc} results in at least one step of balanced wandering. In the worst case, every state must be known before algorithm 1 returns an exploitation policy π_{xplt} , i.e., we must

observe $S\text{Am}_{\text{knw}}$ balanced wandering steps before we can exploit. The following lemma bounds the number of attempted explorations required to have fewer than $S\text{Am}_{\text{knw}}$ balanced wandering steps with at most $\delta/2$ probability.

Lemma A.5 (Number of Exploration Attempts). *For a given MDP $\mathcal{M} = \langle S, A, \mathbf{P}, \mathbf{R}, \gamma \rangle$, confidence level δ , welfare-optimality error ε , horizon T , the probability of having fewer than $S\text{Am}_{\text{knw}}$ balanced wandering steps is less than $\delta/2$ if the number of (T -step) attempted exploration is*

$$M_A = \frac{2pk + f + \sqrt{4pk + f^2}}{2p^2} , \quad (25)$$

where $p = E - \alpha T = \frac{E}{2}$, $k = S\text{Am}_{\text{knw}}$, and $f = \frac{\ln(\frac{2}{\delta})}{2}$.

Proof. Recall that each exploration attempt is a Bernoulli random variable with a probability of success at least $\frac{E}{2}$ as derived in lemma A.4. Thus, we can use the additive-Chernoff to compute the number of attempted explorations required to have fewer than $S\text{Am}_{\text{knw}}$ balanced wandering steps in the worst-case, where every state must be known.

Let $X_1, X_2, X_3, \dots, X_n$ denote independent Bernoulli random variables that take value 1 with probability p and 0 otherwise. Suppose that $X = \sum_{i=1}^n X_i$, then

$$\mathbb{P}[X - \mathbf{E}[X] \geq -\sqrt{nh}] \leq e^{-2h^2} . \quad (26)$$

Suppose that $k = S\text{Am}_{\text{knw}}$. In our setting, $p \geq E - \alpha T = \frac{E}{2}$, $n = M_A$ and $\mathbf{E}[X] = pn = (\frac{E}{2}) M_A$ and $k = pn - \sqrt{nh}$. Therefore, $h = \frac{pn-k}{\sqrt{n}}$.

Setting $e^{-2h^2} = \delta/2$, we get

$$p^2 n^2 + k^2 - 2pnk - nf = 0 , \quad (27)$$

where $f = \frac{1}{2} \ln(\frac{2}{\delta})$. Solving the above equation, we get

$$M_A = n = \frac{2pk + f + \sqrt{4pkf + f^2}}{2p^2} . \quad (28)$$

We note that in order to bound the total number of exploration actions in algorithm 1, m_{knw} must be computed using lemma 4.3 with confidence level $\frac{\delta}{2}$ since we only allocate $\frac{\delta}{2}$ failure probability for the approximation of next-state transition probabilities and rewards. The rest of the failure probability $\frac{\delta}{2}$ is assigned to the failure in exploration attempts. \square

A.2 Proof of Theorem 4.4

Theorem 4.4 (E^4 is KWIK-AF). *Algorithm 1 is a KWIK-AF learner that learns \mathcal{M} w.r.t. the class of all $\lambda\|\cdot\|_\infty$ Lipschitz welfare functions, with sample complexity*

$$m(\varepsilon, \delta, |\mathcal{S}|, |\mathcal{A}|, \gamma, R_{\max}, g) \in \text{Poly} \left(|\mathcal{S}|, |\mathcal{A}|, \log g, R_{\max}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\delta}, \frac{1}{1-\gamma} \right) .$$

Proof. Recall that an algorithm is a KWIK-AF learner if the following two conditions are satisfied with high confidence $1 - \delta$: (A) The number of exploration actions is polynomially bounded, and (B) all the exploitation policies are ε -welfare-optimal.

We now verify that the number of exploration actions in algorithm 1 is polynomially bounded.

The total number of exploration actions is given by the product of number of exploration attempts made by the algorithm (M_A) and the horizon of the temporal escape policy (T). Note that in the worst case, every state must be known before algorithm 1 returns an exploitation policy π_{xplt} . From lemma A.5, we know that for a state to be *known*, each action in that state must be observed m_{knw}

times. Thus, we require SAm_{knw} balanced wandering steps in the worst case. Lemma A.5 bounds the number of exploration attempts (M_A), such that the probability of having fewer than SAm_{knw} balanced wandering steps is less than $\frac{\delta}{2}$. Note that this bound is polynomial in E , α , m_{knw} , and T . Furthermore, from lemma A.4 and lemma 4.3, we know that the values of E , α , m_{knw} and T themselves are polynomial in ε , δ , $|\mathcal{S}|$, $|\mathcal{A}|$, γ , R_{\max} and g . Thus, we confirm that the number of exploration attempts is polynomially bounded with probability at least $1 - \frac{\delta}{2}$.

Next, we consider the second condition of the KWIK-AF framework, i.e., all exploitation policies must be ε -welfare-optimal. Recall that algorithm 1 returns an exploitation policy only when the escape probability of the escape policy is less than the threshold E . In lemma A.4, we show that as long as $\hat{\mathcal{M}}(\alpha, \beta)$ approximates \mathcal{M} over $\mathcal{S} \setminus \mathcal{S}_{\text{unk}}$, and $\frac{6\lambda R_{\max}}{\varepsilon(1-\gamma)} > 1$ and $E \in (0, \frac{1}{2})$ holds, all exploitation policies returned by algorithm 1 are ε -welfare-optimal. According to lemma 4.3, the assumption $\hat{\mathcal{M}}(\alpha, \beta)$ approximates \mathcal{M} over $\mathcal{S} \setminus \mathcal{S}_{\text{unk}}$ fails with probability at most $\frac{\delta}{2}$ when the number of per-state-action pair sample size (m_{knw}) is set to $\left\lceil \ln \left(\frac{2\mathcal{S}\|\mathcal{A}\|(2^{|\mathcal{S}|-2}+2g)}{\delta} \right) \max \left(\frac{1}{2\alpha^2}, \frac{R_{\max}^2}{2\beta^2} \right) \right\rceil$. Note that this is the value of m_{knw} used in algorithm 1. Thus, we confirm that all exploitation policies returned by algorithm 1 are ε -welfare-optimal with probability at least $1 - \frac{\delta}{2}$.

Applying union bound over failure probabilities of both conditions (A) and (B), we obtain that both conditions are satisfied with probability $1 - \delta$ over the course of the E^4 algorithm. Therefore, E^4 is a KWIK-AF learner.

□

We show theorem 3.4.

Theorem 3.4 (Policy-KWIK and PAC-MDP Learners). *Every policy-KWIK learner that outputs deterministic policies is a PAC-MDP learner, in the sense that executing an exploitation policy or exploration action at each timestep produces no more than $m(\varepsilon, \delta, |\mathcal{S}|, |\mathcal{A}|, \gamma, R_{\max})$ total mistakes (i.e., ε -suboptimal actions) with probability at least $1 - \delta$.*

Proof. We proceed by accounting for all timesteps t , and bounding the number of such steps at which an ε -suboptimal action may be taken. The output topology of policy-KWIK and PAC-MDP algorithms differ slightly, as policy-KWIK to learners sometimes output policies instead of actions, however, we bridge this gap by choosing $a_t \leftarrow \pi_t(s_t)$ to select the action our constructed PAC-MDP algorithm will take, i.e., we convert a policy π_t to the action a_t taken at the current state s_t by π_t . In particular, at each successful exploitation step, we show that the action a_t is ε -optimal, and furthermore, while each exploration step may be arbitrarily suboptimal, we need only show that the number of exploration steps remains bounded.

Note that by definition 3.3, with probability at least $1 - \delta$, the total number of exploration actions taken by a KWIK-AF learner over all time does not exceed $m(\varepsilon, \delta, |\mathcal{S}|, |\mathcal{A}|, \gamma, R_{\max})$. We thus conclude that with probability at least $1 - \delta$, no more than $m(\varepsilon, \delta, |\mathcal{S}|, |\mathcal{A}|, \gamma, R_{\max})$ exploration steps occur, each of which may be a mistake. Furthermore, subject to this same probabilistic event, all exploitation policies will be ε -optimal.

Now, to show the PAC-MDP mistake bound, it suffices to show that outputting a near-optimal (deterministic) policy π is sufficient to output a near-optimal action (in the PAC-MDP sense) at s , in particular $\pi(s)$. In service of this result, we now introduce the *state-action value function*

$$Q^\pi(s, a) \doteq \mathbb{E}_{\substack{a_t \sim \pi(s_t) \\ s_{t+1} \sim \mathbf{P}(s_t, a_t, \cdot)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_0 = a \right] = \mathbb{E}_{s_1 \sim \mathbf{P}(s, a, \cdot)} \left[R(s, \pi(s)) + \gamma V^\pi(s_1) \right].$$

Now, observe that

$$\underbrace{\max_a Q^*(s_t, a)}_{\text{OPTIMAL}} - \varepsilon \leq \underbrace{Q^{\pi_t}(s_t, \pi_t(s_t))}_{\text{KWIK MDP}} = Q^{\pi_t}(s_t, a_t) \leq \underbrace{Q^*(s_t, a_t)}_{\text{PAC-MDP}}.$$

Note that $\pi_t(s_t)$ is deterministic given s_t by assumption, thus taking any such action is ε -optimal, and thus not a mistake. \square

B On Fair Planning

B.1 Proof of proposition 4.1

Proposition 4.1 (Welfare-Optimal Planning). *For a concave welfare function $W(\cdot)$, the welfare-optimal policy $\pi^* = \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} W(\mathbf{V}^\pi(s))$ can be computed by first solving*

$$\begin{aligned} \mathbf{d}^* = \operatorname{argmax}_{\mathbf{d} \in \mathbb{R}_{\geq 0}^{\mathcal{S} \times \mathcal{A}}} W \left(\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_1(s, a), \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}(s, a) \mathbf{R}_2(s, a), \dots, \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_g(s, a) \right) \quad (3) \\ \text{such that } \forall s \in \mathcal{S} : \sum_{a \in \mathcal{A}} \mathbf{d}_{s,a} = \mathbf{p}_0(s) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} P(s', a', s) \mathbf{d}_{s',a'} , \end{aligned}$$

and then setting $\pi^*(s, a) = \frac{\mathbf{d}_{s,a}^*}{\sum_{a' \in \mathcal{A}} \mathbf{d}_{s,a'}^*}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Proof. For any policy π , the objective $W\left(\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a}^\pi \mathbf{R}_1(s, a), \dots, \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a}^\pi \mathbf{R}_g(s, a)\right)$ is equivalent to the objective $W(\mathbf{V}^\pi)$ (Wang et al., 2008). Furthermore, there is a one-to-one correspondence between any policy π and the state action occupancy frequency \mathbf{d}^π , and thus, optimizing policy π to maximize the group-welfare objective is equivalent to optimizing occupancy frequency \mathbf{d} to maximize the same objective. The optimization problem in (3) simply follows from using these results along with the Bellman equation of the state-action occupancy measure in (2).

We note that the above optimization problem is concave in state-action occupancy measure \mathbf{d} . This is because the objective in eq. (3) is a concave function of an affine transformation of the state-action occupancy measure \mathbf{d} , making it concave as well; moreover, the constraints are also linear with respect to \mathbf{d} . Thus, we can solve the above optimization problem in polynomial time using any convex-optimization algorithms like subgradient-ascent. We also note that the optimization problem in (3) is similar in structure to the optimization problem for general convex MDPs (Zahavy et al., 2021). Thus, we can alternatively leverage the algorithms proposed by Zahavy et al. (2021) to efficiently solve the above optimization problem. \square

We now show how to compute the welfare-optimal policies for the Nash welfare, Gini index, and Egalitarian welfare objectives by posing them as instances of the problem in (3).

B.1.1 Planning for Egalitarian Welfare

The egalitarian welfare problem is given by

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} \min_{i \in 1, \dots, g} \mathbf{V}_i^\pi(s) .$$

The above problem is equivalent to solving

$$\begin{aligned} \mathbf{d}^* = \operatorname{argmax}_{\mathbf{d} \in \mathbb{R}_{\geq 0}^{\mathcal{S} \times \mathcal{A}}} \min_{i \in 1, \dots, g} \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_i(s, a) \\ \text{s.t. } \sum_{a \in \mathcal{A}} \mathbf{d}_{s,a} = \mathbf{p}_0(s) + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} P(s', a', s) \mathbf{d}_{s',a'}, \forall s \in \mathcal{S} . \end{aligned} \quad (29)$$

Using simple algebraic manipulations, we can further simplify the problem in (29) as

$$\begin{aligned}
\mathbf{d}^* = \operatorname{argmax}_{\mathbf{d} \in \mathbb{R}_{\geq 0}^{\mathcal{S} \times \mathcal{A}}, t \in \mathbb{R}} t \\
\text{s.t. } t \leq \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_i(s, a), \forall i \in 1, \dots, g \\
\sum_{a \in \mathcal{A}} \mathbf{d}_{s,a} = \mathbf{p}_0(s) + \gamma \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} \mathbf{P}(s', a, s) \mathbf{d}_{s',a}, \forall s \in \mathcal{S} .
\end{aligned} \tag{30}$$

We can derive π^* from the optimal state-action occupancy measure \mathbf{d}^* as shown in proposition 4.1.

B.1.2 Planning for Gini welfare

The Gini welfare optimization problem is given by

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} \sum_{i \in 1, \dots, g} \mathbf{w}_i \cdot \mathbf{V}_i^\pi(s) .$$

The above problem is equivalent to solving

$$\begin{aligned}
\mathbf{d}^* = \operatorname{argmax}_{\mathbf{d} \in \mathbb{R}_{\geq 0}^{\mathcal{S} \times \mathcal{A}}} \sum_{i \in 1, \dots, g} \mathbf{w}_i \cdot \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_i(s, a) \\
\text{s.t. } \sum_{a \in \mathcal{A}} \mathbf{d}_{s,a} = \mathbf{p}_0(s) + \gamma \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} \mathbf{P}(s', a, s) \mathbf{d}_{s',a}, \forall s \in \mathcal{S} .
\end{aligned} \tag{31}$$

We apply the procedure described in proposition 4.1 to obtain the optimal policy π^* from the optimal state-action occupancy measure \mathbf{d}^* .

B.1.3 Planning for Nash welfare

The Nash welfare optimization problem is given by

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi_{\mathcal{M}}} \sqrt[p]{\frac{1}{g} \prod_{i \in 1, \dots, g} \mathbf{V}_i^\pi(s)^p} .$$

The above problem is equivalent to solving

$$\begin{aligned}
\mathbf{d}^* = \operatorname{argmax}_{\mathbf{d} \in \mathbb{R}_{\geq 0}^{\mathcal{S} \times \mathcal{A}}} \sum_{i \in 1, \dots, g} \ln \left(\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mathbf{d}_{s,a} \mathbf{R}_i(s, a) \right) \\
\sum_{a \in \mathcal{A}} \text{s.t. } \mathbf{d}_{s,a} = \mathbf{p}_0(s) + \gamma \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} \mathbf{P}(s', a, s) \mathbf{d}_{s',a}, \forall s \in \mathcal{S} .
\end{aligned} \tag{32}$$

We apply the procedure in proposition 4.1 to obtain the optimal policy π^* from the optimal state-action occupancy measure \mathbf{d}^* .