A. Omitted Proofs

A.1. Omitted Proofs for Section 2

Proof of Lemma 1. Let $\hat{\mu}^{\pi}_{T}$ denote the distribution of π on states of M after following π for T steps starting from s. Then we know

$$\begin{aligned}\mathbb{E}_{s\sim\mu^{\pi}} V^{\pi}_{M}(s) &- \frac{1}{T}\mathbb{E}\sum_{t=1}^{T} V^{\pi}_{M}(s_t) \\ &= \sum_{i=1}^{n} \left(\mu^{\pi}(s_i) - \hat{\mu}^{\pi}_{T}(s_i)\right) V^{\pi}_{M}(s_i) \\ &\le \sum_{i=1}^{n} \left|\mu^{\pi}(s_i) - \hat{\mu}^{\pi}_{T}(s_i)\right| V^{\pi}_{M}(s_i) \\ &\le \frac{\epsilon}{1-\gamma}.\end{aligned}$$

The last inequality is due to the following observations: (i) $V^{\pi}_{M}(s_i) \le \frac{1}{1-\gamma}$ as rewards are in $[0,1]$ and (ii) $\Sigma_{i=1}^{n} |\mu^{\pi}(s_i) - \hat{\mu}^{\pi}_{T}(s_i)| \le \epsilon$ since T is at least the ϵ-mixing time of π. $\square$

A.2. Omitted Proofs for Section 3

We first state the following useful Lemma about M.

Lemma 11. *Let M be the MDP in Definition 6. Then for any $i \in \{1,\ldots,n\}$, $V^{*}_{M}(s_i) < \frac{1+2\gamma^{n-i+1}}{2(1-\gamma)}$.*

Proof.

$$\begin{aligned}V^{*}_{M}(s_i) &= \text{discounted reward before reaching state } n \\ &\quad + \text{discounted reward from staying at state } n \\ &< \left[\sum_{t=1}^{n-i-1} \frac{\gamma^t}{2}\right] + \frac{\gamma^{n-i+1}}{1-\gamma} \\ &= \left[\frac{1}{2}\left(\frac{1}{1-\gamma} - \frac{\gamma^{n-i}}{1-\gamma}\right)\right] + \frac{\gamma^{n-i+1}}{1-\gamma} \\ &= \frac{1-\gamma^{n-i}}{2(1-\gamma)} + \frac{\gamma^{n-i+1}}{1-\gamma} \\ &= \frac{1+\gamma^{n-i}(2\gamma-1)}{2(1-\gamma)} \\ &< \frac{1+2\gamma^{n-i+1}}{2(1-\gamma)},\end{aligned}$$

via two applications of the summation formula for geometric series. $\square$

Proof of Theorem 3. We prove Theorem 3 for the special case of $k = 2$ first. Consider coupling the run of a fair algorithm $\mathcal{L}$ on both $M(0.5)$ and $M(1)$. To achieve this, we can fix the randomness of $\mathcal{L}$ up front, and use the same randomness on both MDPs. The set of observations and hence the actions taken on both MDPs are identical until $\mathcal{L}$ reaches state s_n. Until then, with probability at least $1-\delta$, $\mathcal{L}$ must play L and R with equal probability in order to satisfy fairness (since, for $M(0.5)$, the only fair policy is to play both actions with equal probability at each time step). We will upper-bound the optimality of uniform play and lower-bound the number of rounds before which s_n is visited by uniformly random play.

Let $f_\gamma = \lceil \frac{1}{1-\sqrt[3]{\gamma}} \rceil$ and $\mathcal{T} = 2^{n-2f_\gamma}$ for $n \ge 100(f_\gamma)^2$. First observe that the probability of reaching a fixed state s_i for any $i \ge n - f_\gamma$ from a random walk of length $\mathcal{T}$ is upper bounded by the probability that the random walk takes $i \ge n - f_\gamma$ consecutive steps to the right in the first $\mathcal{T}$ steps. This probability is at most $p = 2^{n-2f_\gamma}(\frac{1}{2})^{n-f_\gamma} = 2^{-f_\gamma}$ for any fixed i. Since reaching any state $i > i'$ requires reaching state i', the probability that the $\mathcal{T}$ step random walk arrives in any state s_i for $i \ge n - f_\gamma$ is also upper bounded by p.

Next, we observe that $V^{*}_{M}(s_i)$ is a nondecreasing function of i for both MDPs. Then the average V^{*}_{M} values of the visited states of *any* fair policy can be broken into two pieces: the average conditioned on (the probability at least $1-\delta$ event) that the algorithm plays uniformly at random before reaching state s_n *and* never reaching a state beyond s_{n-f_γ}, and the average conditioned on (the probability at most δ event) that the algorithm does not make uniformly random choices *or* the uniform random walk of length $\mathcal{T}$ reaches a state beyond s_{n-f_γ}. So, we have that

$$\begin{aligned}\frac{1}{\mathcal{T}}\mathbb{E}\sum_{t=1}^{\mathcal{T}} V^{*}_{M}(s_t) &\le (1-p-\delta)\, V^{*}_{M}(s_{n-f_\gamma}) + (p+\delta)\,\frac{1}{1-\gamma} \\ &\le (1-p-\delta)\,\frac{1+2\gamma^{f_\gamma+1}}{2(1-\gamma)} + (p+\delta)\,\frac{1}{1-\gamma}.\end{aligned}$$

The first inequality follows from the fact that $V^{*}_{M}(s_i) \le \frac{1}{1-\gamma}$ for all i, and the second from Lemma 11 along with V^{*}_{M} values being nondecreasing in i. Putting it all together,

$$\begin{aligned}&\mathbb{E}_{s\sim\mu^{*}}\ V^{*}_{M}(s) - \frac{1}{\mathcal{T}}\mathbb{E}\sum_{t=1}^{\mathcal{T}} V^{*}_{M}(s_t) \\ &\quad\ge \frac{1}{1-\gamma} - \left[(1-p-\delta)\,\frac{1+2\gamma^{f_\gamma+1}}{2(1-\gamma)} + (p+\delta)\,\frac{1}{1-\gamma}\right] \\ &\quad= \frac{1-p-\delta}{1-\gamma}\left[1 - \frac{1+2\gamma^{f_\gamma+1}}{2}\right].\end{aligned}$$

So ϵ-optimality requires

$$\frac{2\epsilon}{1-\gamma} \ge \frac{1-p-\delta}{1-\gamma}\left[1 - \frac{1+2\gamma^{f_\gamma+1}}{2}\right]. \tag{4}$$

However, if $\epsilon < \frac{1}{8}$ we get

$$\frac{2\epsilon}{1-\gamma} < \frac{1-0.04-1/4}{1-\gamma}\left[1-\frac{1+2\times e^{-3}}{2}\right] < \frac{1-2^{-f_\gamma}-\delta}{1-\gamma}\left[1-\frac{1+2\gamma^{f_\gamma+1}}{2}\right],$$

where the third inequality follows when $\delta < \frac{1}{4}$ and $\gamma > \frac{1}{2}$. This means $\epsilon < \frac{1}{8}$ makes ϵ-optimality impossible, as desired.

Throughout we considered the special case of $k = 2$ and proved a lower bound of $\Omega(2^n)$ time steps for any fair algorithm satisfying the ϵ-optimality condition. However, it is easy to see that MDP M in Definition 6 can be easily modified in a way that $k-1$ of the actions from state s_i reach state s_1 and only one action in each state s_i reaches states $s_{\min\{i+1,n\}}$. Hence, a lower bound of $\Omega(k^n)$ time steps can be similarly proved. □

Proof of Theorem 4. We mimic the argument used to prove Theorem 3 with the difference that, until visiting s_n, $\mathcal{L}$ may not play R with probability more than $\frac{1}{2}+\alpha$ (as opposed to $\frac{1}{2}$ in Theorem 3). Let $f_\gamma = \lceil \frac{1}{1-\sqrt[3]{\gamma}} \rceil$ and $\mathcal{T} = (\frac{2}{1+2\alpha})^{n-2f_\gamma}$ for $n \geq 100(f_\gamma)^2$. By a similar process as in Theorem 3, the probability of reaching state s_i for any $i \geq n - f_\gamma$ from a random walk of length $\mathcal{T}$ is bounded by $p = (\frac{2}{1+2\alpha})^{-f_\gamma}$, and so the probability that the $\mathcal{T}$ steps random walk arrives in any state s_i for $i \geq n - f_\gamma$ is bounded by p. Carrying out the same process used to prove Theorem 3 then once more implies that ϵ-optimality requires Equation 4 to hold when $\delta < \frac{1}{4}$, $\alpha < \frac{1}{4}$ and $\gamma > \frac{1}{2}$. Hence, $\epsilon < \frac{1}{8}$ violates this condition as desired.

Finally, throughout we considered the special case of $k = 2$. The same trick as in the proof of Theorem 3 can be used to prove the lower bound of $\Omega((\frac{k}{1+k\alpha})^n)$ time steps for any fair algorithm satisfying the ϵ-optimality condition. □

Proof of Theorem 5. We also prove Theorem 5 for the special case of $k = 2$ first, again considering the MDP in Definition 6. We set the size of the state space in M to be $n = \lceil \frac{\log(\frac{1}{2\alpha})}{1-\gamma} \rceil$. Then given the parameter ranges, for any i, $Q^*_M(s_i, R) - Q^*_M(s_i, L) > \alpha$ in M(1). Therefore, any approximate-action fair algorithm should play actions R and L with equal probability.

Let $\mathcal{T} = 2^{cn} = \Omega((2^{1/(1-\gamma)})^c)$. First observe that the probability of reaching a fixed state s_i for any $i \geq (c+1)n/2$ from a random walk of length $\mathcal{T}$ is upper bounded by the probability that the random walk takes $i \geq (c+1)n/2$ consecutive steps to the right in the first $\mathcal{T}$ steps. This probability is at most $p = 2^{cn}2^{-(c+1)n/2} = 2^{(c-1)n/2}$ for any fixed i. Then the probability that the $\mathcal{T}$ steps random walk arrives in any state s_i for $i \geq (c+1)n/2$ is also upper bounded by p.

Next, we observe that $V^*_M(s_i)$ is a nondecreasing function of i, for both MDPs. Then the average V^*_M values of the visited states of *any* fair policy can be broken into two pieces: the average conditioned on the $1-\delta$ fairness *and* never reaching a state beyond $s_{(c+1)n/2}$, and the average when fairness might be violated *or* the uniform random walk of length $\mathcal{T}$ reaches a state beyond $s_{(c+1)n/2}$. So, we have that

$$\begin{aligned}\frac{1}{\mathcal{T}}\mathbb{E}\sum_{t=1}^{\mathcal{T}} V^*_M(s_t) \leq\ & (1-p-\delta)\, V^*_M(s_{(c+1)n/2}) \\ &+ (p+\delta)\,\frac{1}{1-\gamma} \\ &\leq (1-p-\delta)\,\frac{1+(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}}{2(1-\gamma)} \\ &= (p+\delta)\,\frac{1}{1-\gamma}.\end{aligned}$$

The first inequality follows from the fact that $V^*_M(s_i) \leq \frac{1}{1-\gamma}$ for all i, and the second from (the line before the last in) Lemma 11 along with V^*_M values being nondecreasing in i. Putting it all together,

$$\begin{aligned}&\mathbb{E}_{s\sim\mu^*}\, V^*_M(s) - \frac{1}{\mathcal{T}}\mathbb{E}\sum_{t=1}^{\mathcal{T}} V^*_M(s_t) \\ &\geq \frac{1}{1-\gamma} - (1-p-\delta)\,\frac{1+(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}}{2(1-\gamma)} \\ &- (p+\delta)\,\frac{1}{1-\gamma} \\ &= \frac{1-p-\delta}{1-\gamma}\left[1-\frac{1+(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}}{2}\right] \\ &= \frac{1-p-\delta}{1-\gamma}\left[\frac{1}{2}-\frac{(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}}{2}\right].\end{aligned}$$

So ϵ-optimality requires

$$\frac{2\epsilon}{1-\gamma} \geq \frac{1-p-\delta}{1-\gamma}\left[\frac{1}{2}-\frac{(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}}{2}\right].$$

Rearranging and using $\delta < \frac{1}{4}$, we get that ϵ-optimality requires

$$4\epsilon \geq \left[0.75 - 2^{\frac{(c-1)n}{2}}\right]\left[1-(2\gamma-1)\gamma^{\frac{(1-c)n}{2}}\right]$$

and expand n to get

$$\begin{aligned}\epsilon \geq \frac{1}{4}&\left[0.75 - 2^{\frac{(c-1)\log(\frac{1}{2\alpha})}{2(1-\gamma)}}\right]\times \\ &\left[1-(2\gamma-1)\gamma^{\frac{(1-c)\log(\frac{1}{2\alpha})}{2(1-\gamma)}}\right] \equiv \frac{xy}{4}.\end{aligned}$$

Noting that x is minimized when $2^{\frac{(c-1)\log(\frac{1}{2\alpha})}{2(1-\gamma)}}$ is maximized, and that this quantity is maximized when $\frac{\log(\frac{1}{2\alpha})}{2(1-\gamma)}$ is minimized (as $c-1$ is negative), we get that ϵ-optimality requires

$$\epsilon \geq \frac{\left[0.75 - 2^{\frac{c-1}{1-\gamma}}\right] y}{4}$$

from $\alpha < \frac{1}{8}$. Similarly, $\alpha < \frac{1}{8}$ implies that ϵ-optimality requires

$$\epsilon \geq \frac{\left[0.75 - 2^{\frac{c-1}{1-\gamma}}\right]\left[1-(2\gamma-1)\gamma^{\frac{1-c}{1-\gamma}}\right]}{4}.$$

Note that $0.75 - 2^{\frac{c-1}{1-\gamma}}$ is minimized when γ is small, so $\gamma > c$ implies that ϵ-optimality requires

$$\begin{aligned}\epsilon &\geq \frac{\left[0.75 - 2^{-1}\right]\left[1-(2\gamma-1)\gamma^{\frac{1-c}{1-\gamma}}\right]}{4} \\ &\geq \frac{1}{16}\left[1-(2\gamma-1)\gamma^{\frac{1-c}{2(1-\gamma)}}\right].\end{aligned}$$

Conversely, $1-(2\gamma-1)\gamma^{\frac{1-c}{1-\gamma}}$ is minimized when γ is large, so as

$$\lim_{\gamma\to 1}(2\gamma-1)\gamma^{\frac{1-c}{1-\gamma}} = e^{c-1}$$

we get that ϵ-optimality requires

$$\epsilon \geq \frac{1}{16}\left(1-e^{c-1}\right).$$

Finally, the same trick as in the proof of Theorem 3 can be used to prove the $\Omega((k^{1/(1-\gamma)})^c)$ lower bound for $k > 2$ actions. □

A.3. Omitted Proofs for Section 4

Proof of Lemma 8. We first show that either

- there exists an *exploitation policy* π in M_Γ such that

$$\frac{1}{T}\max_{\bar{\pi}\in\Pi}\mathbb{E}\sum_{t=1}^{T}V_M^{\bar{\pi}}\left(\bar{\pi}^t(s),T\right) - \frac{1}{T}\mathbb{E}\sum_{t=1}^{T}V_{M_\Gamma}^{\pi}\left(\pi^t(s),T\right) \leq \beta$$

  where the random variables $\pi^t(s)$ and $\bar{\pi}^t(s)$ denote the states reached from s after following π and $\bar{\pi}$ for t steps, respectively, or
- there exists an *exploration policy* π in M_Γ such that the probability that a walk of $2T$ steps from s following π will terminate in s_0 exceeds $\frac{\beta}{T}$.

Let π be a policy in M satisfying

$$\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}V_M^{\pi}(\pi^t(s),T) = \frac{1}{T}\max_{\bar{\pi}\in\Pi}\mathbb{E}\sum_{t=1}^{T}V_M^{\pi'}(\bar{\pi}^t(s),T) := \tilde{V}.$$

For any state s', let $p(s')$ denote all the paths of length T in M that start in s', $q(s')$ denote all the paths of length T in M that start in s' such that all the states in every path of length T in $q(s')$ are in Γ and $r(s')$ all the paths of length T in M that start in s' such that at least one state in every path of length T in $r(s')$ is not in Γ. Suppose

$$\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}V_{M_\Gamma}^{\pi}(\pi^t(s)) < \tilde{V} - \beta.$$

Otherwise, π already witnesses the claim. We show that a walk of $2T$ steps from s following π will terminate in s_0 with probability of at least $\frac{\beta}{T}$. First,

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}V_M^{\pi}(\pi^t(s),T) &= E\sum_{t=1}^{T}\sum_{p(\pi^t(s))}\mathbb{P}[p(\pi^t(s))]V_M(p(\pi^t(s))) \\ &= \mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}[q(\pi^t(s))]V_M(q(\pi^t(s))) \\ &+ \mathbb{E}\sum_{t=1}^{T}\sum_{r(\pi^t(s))}\mathbb{P}[r(\pi^t(s))]V_M(r(\pi^t(s)))\end{aligned}$$

since $p(\pi^t(s)) = q(\pi^t(s)) \cup r(\pi^t(s))$, which is a disjoint union. Next,

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}[q(\pi^t(s))]V_M(q(\pi^t(s))) \\ &= \mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}_{M_\Gamma}^{\pi}[q(\pi^t(s))]V_{M_\Gamma}(q(\pi^t(s))) \\ &\leq \mathbb{E}\sum_{t=1}^{T}V_{M_\Gamma}^{\pi}(\pi^t(s),T),\end{aligned}$$

where the equality is due to Definition 9 and the definition of q, and the inequality follows because $V_{M_\Gamma}^{\pi}(\pi^t(s),T)$ is the sum over all the T-paths in M_Γ, not just those that avoid the absorbing state s_0. Therefore by our original assumption on π,

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}[q(\pi^t(s))]V_M(q(\pi^t(s))) \\ &\leq \mathbb{E}\sum_{t=1}^{T}V_{M_\Gamma}^{\pi}(\pi^t(s),T) < T\tilde{V} - T\beta.\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\sum_{r(\pi^t(s))}\mathbb{P}[r(\pi^t(s))]V_M(r(\pi^t(s)))\\
&=\mathbb{E}\sum_{t=1}^{T}V_M^{\pi}(\pi^t(s),T)\\
&-\mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}[q(\pi^t(s))]V_M(q(\pi^t(s)))\\
&=T\tilde{V}-\mathbb{E}\sum_{t=1}^{T}\sum_{q(\pi^t(s))}\mathbb{P}[q(\pi^t(s))]V_M(q(\pi^t(s)))\geq T\beta,
\end{aligned}$$

where the last step is the result of applying the previous inequality. However,

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\sum_{r(\pi^t(s))}\mathbb{P}[r(\pi^t(s))]V_M(r(\pi^t(s)))\\
&\leq T\mathbb{E}\sum_{t=1}^{T}\sum_{r(\pi^t(s))}\mathbb{P}[r(\pi^t(s))],
\end{aligned}$$

because it is immediate that $V_M(r(\pi^t(s))) \leq T$ for all $\pi^t(s)$. So $T\beta \leq T\mathbb{E}\sum_{t=1}^{T}\sum_{r(\pi^t(s))}\mathbb{P}[r(\pi^t(s))]$. Finally, if we let $\mathbb{P}_{2T}^{\pi}$ denote the probability that a walk of $2T$ steps following π terminates in s_0, i.e. the probability that π escapes to an unknown state within $2T$ steps, then for each $t \in [T]$, $\mathbb{E}\sum_{r(\pi^t(s))} \leq T\mathbb{P}_{2T}^{\pi}$. It follows that

$$T\beta \leq T^2\mathbb{P}_{2T}^{\pi}$$

and rearranging yields $\mathbb{P}_{2T}^{\pi} \geq \frac{\beta}{T}$ as desired.

Next, note that the exploitation policy (if it exists) can be derived by computing the optimal policy in M_Γ. Moreover, the exploration policy (if it exists) in the exploitation MDP M_Γ can indeed be derived by computing the optimal policy in the exploration MDP $M_{[n]\setminus\Gamma}$ as observed by (Kearns and Singh, 2002). Finally, by Observation 5, any optimal policy in $\hat{M}_\Gamma^\alpha$ ($\hat{M}_{[n]\setminus\Gamma}^\alpha$) is an optimal policy in $\hat{M}_\Gamma$ ($\hat{M}_{[n]\setminus\Gamma}$) □

To prove Lemma 10, we need some useful background adapted from Kearns and Singh (2002).

Definition 8 (Definition 7, Kearns and Singh (2002))**.** *Let M and $\hat{M}$ be two MDPs with the same set of states and actions. We say $\hat{M}$ is a β-approximation of M if*

- *For any state s,*

$$\bar{R}_M(s)-\beta \leq \bar{R}_{\hat{M}}(s) \leq \bar{R}_M(s)+\beta.$$

- *For any states s and s' and action a,*

$$P_M(s,a,s')-\beta \leq P_{\hat{M}}(s,a,s') \leq P_M(s,a,s')+\beta.$$

Lemma 12 (Lemma 5, Kearns and Singh (2002))**.** *Let M be an MDP and Γ the set of known states of M. For any $s, s' \in \Gamma$ and action $a \in A$, let $\hat{P}_M(s,a,s')$ denote the empirical probability transition estimates obtained from the visits to s. Moreover, for any state $s \in \Gamma$ let $\hat{\bar{R}}(s)$ denote the empirical estimates of the average reward obtained from visits to s. Then with probability at least $1-\delta$,*

$$|\hat{P}_M(s,a,s')-P_M(s,a,s')| = O\left(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}}\right),$$

and

$$|\hat{\bar{R}}_M(s)-\bar{R}_M(s)| = O\left(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}}\right).$$

Lemma 12 shows that $\hat{M}_\Gamma$ and $\hat{M}_{[n]\setminus\Gamma}$ are $O(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}})$-approximation MDPs for M_Γ and $M_{[n]\setminus\Gamma}$, respectively.

Lemma 13 (Lemma 4, Kearns and Singh (2002))**.** *Let M be an MDP and $\hat{M}$ its $O(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}})$-approximation. Then for any policy $\pi \in \Pi$ and any state s and action a*

$$V_M^\pi(s)-\min\{\epsilon,\alpha\} \leq V_{\hat{M}}^\pi(s) \leq V_M^\pi(s)+\min\{\epsilon,\frac{\alpha}{4}\},$$

and

$$\begin{aligned}
Q_M^\pi(s,a)-\min\{\frac{\alpha}{4},\epsilon\} &\leq Q_{\hat{M}}^\pi(s,a)\\
&\leq Q_M^\pi(s,a)+\min\{\frac{\alpha}{4},\epsilon\}.
\end{aligned}$$

Proof of Lemma 10. By Definition 7 and Lemma 12, $\hat{M}_\Gamma$ is a $O(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}})$-approximation of M_Γ. Then the statement directly follows by applying Lemma 13. □

Rest of the Proof of Theorem 6. The only remaining part of the proof of Theorem 6 is the analysis of the probability of failure of **Fair-E**^{3}. To do so, we break down the probability of failure of **Fair-E**^{3} by considering the following (exhaustive) list of possible failures:

1. At some known state the algorithm has a poor approximation of the next step, causing $\hat{M}_\Gamma$ to not be a $O(\frac{\min\{\epsilon,\alpha\}^2}{n^2 H_\epsilon^{\gamma\,4}})$-approximation of M_Γ.
2. At some known state the algorithm has a poor approximation of the Q_M^* values for one of the actions.
3. Following the exploration policy for $2T_\epsilon^*$ steps fails to yield enough visits to unknown states.
4. At some known state, the approximation value of that state in $\hat{M}_\Gamma$ is not an accurate estimate for the value of the state in M_Γ.

We allocate $\frac{\delta}{4}$ of our total probability of failure to each of these sources:

1. Set $\delta' = \frac{\delta}{4n}$ in Lemma 10.
2. Set $\delta' = \frac{\delta}{4nk}$ in Theorem 7.
3. By Lemma 8, each attempted exploration is a Bernoulli trial with probability of success of at least $\frac{\epsilon}{4T^*_\epsilon}$. In the worst case we might need to make every state known before exploiting, leading to the nm_Q trajectories (m_Q as Equation 3 in Definition 7) of length H^γ_ϵ. Therefore, the probability of taking fewer than nm_Q trajectories of length H^γ_ϵ would be bounded by $\frac{\delta}{4}$ if the number of $2T^*_\epsilon$ steps explorations is at least

$$m_{\text{exp}} = O\left(\frac{T^*_\epsilon nm_Q}{\epsilon}\log\left(\frac{n}{\delta}\right)\right). \tag{5}$$

4. Set $\delta' = \frac{\delta}{4m_{\text{exp}}}$ (m_{exp} as defined in Equation 5) in Lemma 10, as **Fair-E**^{3} might make $2T^*_\epsilon$ steps explorations up to m_{exp} times.

□

A.4. Relaxing Assumption 2

Throughout Sections 4.3 and 4.4 we assumed that T^*_ϵ, the ϵ-mixing time of the optimal policy π^*, was known (see Assumption 2). Although **Fair-E**^{3} uses the knowledge of T^*_ϵ to decide whether to follow the exploration or exploitation policy, Lemma 8 continues to hold even without this assumption. Note that **Fair-E**^{3} is parameterized by T^*_ϵ and for any input T^*_ϵ runs in time **poly**(T^*_ϵ). Thus if T^*_ϵ is unknown, we can simply run **Fair-E**^{3} for $T^*_\epsilon = 1, 2, \ldots$ sequentially and the running time and sample complexity will still be **poly**(T^*_ϵ). Similar to the analysis of **Fair-E**^{3} when T^*_ϵ is known we have to run the new algorithm for sufficiently many steps so that the possibly low V^*_M values of the visited states in the early stages are dominated by the near-optimal V^*_M values of the visited states for large enough guessed values of T^*_ϵ.

B. Observations on Optimality and Fairness

Observation 1. *For any MDP M, there exists an optimal policy π^* such that π^* is fair.*

Proof. In time t, let state s_t denote the state from which π chooses an action. Let $a^* = \text{argmax}_a\, Q^*_M(s_t, a)$ and $A^*(s_t) = \{a \in A \mid Q^*_M(s_t, a) = Q^*_M(s_t, a^*)\}$. The policy of playing an action uniformly at random from $A^*(s_t)$ in state s_t for all t, is fair and optimal. □

Approximate-action fairness, conversely, can be satisfied by *any* optimal policy, even a deterministic one.

Observation 2. *Let π^* be an optimal policy in MDP M. Then π^* is approximate-action fair.*

Proof. Assume that π^* is not approximate-action fair. Given state s, the action that π^* takes from s is uniquely determined since π^* is deterministic we may denote it by a^*. Then there exists a time step in which π^* is in state s and chooses action $a^*(s)$ such that there exists another action a with

$$Q^*_M(s, a) > Q^*_M(s, a^*(s)) + \alpha,$$

a contradiction of the optimality of π^*. □

Observations 1 and 2 state that policies with optimal performance are fair; we now state that playing an action uniformly at random is also fair.

Observation 3. *An algorithm that, in every state, plays each action uniformly at random (regardless of the history) is fair.*

Proof. Let $\mathcal{L}$ denote an algorithm that in every state plays uniformly at random between all available actions. Then $\mathcal{L}(s, h_{t-1})_a = \mathcal{L}(s, h_{t-1})_{a'}$ regardless of state s, (available) action a, or history h_{t-1}. $Q^*_M(s, a) > Q^*_M(s, a') + \alpha \Rightarrow \mathcal{L}(s, h_{t-1})_a \geq \mathcal{L}(s, h_{t-1})_{a'}$ then follows immediately, which guarantees both fairness and approximate-action fairness. □

Observation 4. *Let M be an MDP and M^α the α-restricted MDP of M. Let π be a policy in M^α. Then π is α-action fair.*

Proof. Assume π is not α-action fair. Then there must exist round t, state s, and action a such that $Q^*_M(s, a) > Q^*_M(s, a') + \alpha$ and $\mathcal{L}(s, h_{t-1})_a < \mathcal{L}(s, h_{t-1})_{a'}$. Therefore $\mathcal{L}(s, h_{t-1})_{a'} > 0$, so M^α must include action a' from state s. But this is a contradiction, as in state s M^α only includes actions a' such that $Q^*_M(s, a') + \alpha \geq Q^*_M(s, a)$. π is therefore α-action fair. □

Observation 5. *Let M be an MDP and M^α the α-restricted MDP of M. Let π^* be an optimal policy in M^α. Then π^* is also optimal in M.*

Proof. If π^* is not optimal in M, then there exists a state s and action a such that $Q^*_M(s, a) > \mathbb{E}_{a^*(s)\sim\pi^*(s)} Q^*_M(s, a^*(s))$ where $a^*(s)$ is drawn from $\pi^*(s)$ and the expectation is taken over choices of $a^*(s)$. This is a contradiction because action a is available from state s in M^α by Definition 5. □

C. Omitted Details of Fair-E^3

We first formally define the exploitation MDP M_Γ and the exploration MDP $M_{[n]\setminus\Gamma}$:

Definition 9 (Definition 9, Kearns and Singh (2002)). *Let $M = (\mathcal{S}_M, \mathcal{A}_M, P_M, R_M, T, \gamma)$ be an MDP with state space $\mathcal{S}_M$ and let $\Gamma \subset \mathcal{S}_M$. We define the* exploration MDP $M_\Gamma = (\mathcal{S}_{M_\Gamma}, \mathcal{A}_M, P_{M_\Gamma}, R_{M_\Gamma}, T, \gamma)$ *on Γ where*

- $\mathcal{S}_{M_\Gamma} = \Gamma \cup \{s_0\}$.
- *For any state $s \in \Gamma$, $\bar{R}_{M_\Gamma}(s) = \bar{R}_M(s)$, rewards in M_Γ are deterministic, and $\bar{R}_{M_\Gamma}(s_0) = 0$.*
- *For any action a, $P_{M_\Gamma}(s_0, a, s_0) = 1$. Hence, s_0 is an absorbing state.*
- *For any states $s_1, s_2 \in \Gamma$ and any action a, $P_{M_\Gamma}(s_1, a, s_2) = P_M(s_1, a, s_2)$, i.e. transitions between states in Γ are preserved in M_Γ.*
- *For any state $s_1 \in \Gamma$ and any action a, $P_{M_\Gamma}(s_1, a, s_0) = \Sigma_{s_2 \notin \Gamma} P_M(s_1, a, s_2)$. Therefore, all the transitions between a state in Γ and states not in Γ are directed to s_0 in M_Γ.*

Definition 10 (Implicit, Kearns and Singh (2002)). *Given MDP M and set of known states Γ, the* exploration MDP *$M_{[n]\setminus\Gamma}$ on Γ is identical to the exploitation MDP M_Γ except for its reward function. Specifically, rewards in $M_{[n]\setminus\Gamma}$ are deterministic as in M_Γ, but for any state $s \in \Gamma$, $\bar{R}_{M_{[n]\setminus\Gamma}}(s) = 0$, and $\bar{R}_{M_{[n]\setminus\Gamma}}(s_0) = 1$.*

We next define the approximation MDPs $\hat{M}_\Gamma$ and $\hat{M}_{[n]\setminus\Gamma}$ which are defined over the same set of states and actions as in M_Γ and $M_{[n]\setminus\Gamma}$, respectively.

Let M be an MDP and Γ the set of known states of M. For any $s, s' \in \Gamma$ and action $a \in A$, let $\hat{P}_{M_\Gamma}(s, a, s')$ denote the empirical probability transition estimates obtained from the visits to s. Moreover, for any state $s \in \Gamma$ let $\hat{\bar{R}}_{M_\Gamma}(s)$ denote the empirical estimates of the average reward obtained from visits to s. Then $\hat{M}_\Gamma$ is identical to M_Γ except that:

- in any known state $s \in \Gamma$, $\hat{R}_{\hat{M}_\Gamma}(s) = \hat{\bar{R}}_{M_\Gamma}(s)$.
- for any $s, s' \in \Gamma$ and action $a \in A$, $P_{\hat{M}_\Gamma}(s, a, s') = \hat{P}_{M_\Gamma}(s, a, s')$.

Also $\hat{M}_{[n]\setminus\Gamma}$ is identical to $M_{[n]\setminus\Gamma}$ except that:

- for any $s, s' \in \Gamma$ and action $a \in A$, $P_{\hat{M}_{[n]\setminus\Gamma}}(s, a, s') = \hat{P}_{M_{[n]\setminus\Gamma}}(s, a, s')$.