A. Omitted Proofs

A.1. Omitted Proofs for Section 2

Proof of Lemma 1. Let \( \mu^\pi_T \) denote the distribution of \( \pi \) on states of \( M \) after following \( \pi \) for \( T \) steps starting from \( s \). Then we know

\[
\mathbb{E}_{s \sim \mu^\pi} V_M^\pi(s) - \frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi(s_t) = \sum_{i=1}^{n} (\mu^\pi(s_i) - \mu_T^\pi(s_i)) V_M^\pi(s_i) \leq \frac{\varepsilon}{1 - \gamma}.
\]

The last inequality is due to the following observations: (i) \( V_M^\pi(s_i) \leq \frac{1}{1 - \gamma} \) as rewards are in \([0, 1]\) and (ii) \( \sum_{i=1}^{n} |\mu^\pi(s_i) - \mu_T^\pi(s_i)| \leq \varepsilon \) since \( T \) is at least the \( \epsilon \)-mixing time of \( \pi \).

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

Next, we observe that \( V_M^\pi(s_i) \) is a nondecreasing function of \( i \) for both MDPs. Then the average \( V_M^\pi \), 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 \( s_{n-f_\gamma} \), and the average conditioned on (the probability at most \( \delta \) event) that the algorithm does not make uniformly random choices or the uniform random walk of length \( T \) reaches a state beyond \( s_{n-f_\gamma} \). So, we have that

\[
\frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi(s_t) \leq (1 - p - \delta) V_M^\pi(s_{n-f_\gamma}) + (p + \delta) \frac{1}{1 - \gamma} \\
\leq (1 - p - \delta) \frac{1 + 2^{f_\gamma+1}}{2(1 - \gamma)} + (p + \delta) \frac{1}{1 - \gamma}.
\]

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

\[
\mathbb{E}_{s \sim \mu^\pi} V_M^\pi(s) - \frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi(s_t) \geq \frac{1}{1 - \gamma} - \frac{1}{1 - \gamma} \left[ (1 - p - \delta) \frac{1 + 2^{f_\gamma+1}}{2(1 - \gamma)} + (p + \delta) \frac{1}{1 - \gamma} \right] \\
= \frac{1}{1 - \gamma} \left[ 1 - \frac{1 + 2^{f_\gamma+1}}{2} \right].
\]

So \( \epsilon \)-optimality requires

\[
2\epsilon \geq \frac{1}{1 - \gamma} \left[ 1 - \frac{1 + 2^{f_\gamma+1}}{2} \right].
\]
However, if $\epsilon < \frac{1}{8}$ we get

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

where the third inequality follows when $\delta < \frac{1}{4}$ and $\gamma > \frac{1}{2}$. This means $\epsilon < \frac{1}{8}$ makes $\epsilon$-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 $\epsilon$-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 $s_j$ 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 $s_n$, $L$ may not play $R$ with probability more than $\frac{1}{4} + \alpha$ (as opposed to $\frac{1}{2}$ in Theorem 3). Let $f_y = \left[ \frac{1}{1 - \frac{1}{\sqrt{2}} - \gamma} \right]$ and $T = (\frac{2}{1 + \sqrt{2}})^{n-2f_y}$ for $n \geq 100(f_y)^2$. By a similar process as in Theorem 3, the probability of reaching state $s_i$ for any $i \geq n - f_y$ from a random walk of length $T$ is bounded by $p_{\gamma} = (\frac{2}{1 + \sqrt{2}})^{n-2f_y}$, and so the probability that the $T$ steps random walk arrives in any state $s_i$ for $i \geq n - f_y$ is bounded by $p$. Carrying out the same process used to prove Theorem 3 then once more implies that $\epsilon$-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 $\Omega((\frac{k}{1 + k/\sqrt{n}})^n)$ time steps for any fair algorithm satisfying the $\epsilon$-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 $n = \left\lceil \frac{-\log(1 - \gamma)}{2} \right\rceil$. 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 $T = 2^n = \Omega((2^{1/(1-\gamma)})^n)$. 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 $T$ 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 $T$ steps. This probability is at most $p = 2^{n-2-(c+1)n/2} = 2^{(c-1)n/2}$ for any fixed $i$. Then the probability that the $T$ 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 $T$ reaches a state beyond $s_{(c+1)n/2}$. So, we have that

$$\frac{1}{T} E \sum_{t=1}^{T} V_M^*(s_i) \leq 1 - (p - \delta) V_M^*(s_{(c+1)n/2}) + (p + \delta) \frac{1}{1 - \gamma} \leq (1 - \delta) + \frac{2(\gamma - 1)}{2(1 - \gamma)} \frac{(1 - \gamma)}{2} = (p + \delta) \frac{1}{1 - \gamma}.$$

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,

$$E_{\pi^*} V_M^*(s) - \frac{1}{T} E \sum_{t=1}^{T} V_M^*(s_i) \geq \frac{1}{1 - \gamma} - \frac{(p - \delta)}{1 - \gamma} \frac{1 - (2\gamma - 1)\gamma^{\frac{(1 - \gamma)n}{2}}}{2(1 - \gamma)} = \frac{1}{1 - \gamma}.$$ 

So $\epsilon$-optimality requires

$$\frac{2\epsilon}{1 - \gamma} \geq \frac{1}{2} - \frac{(2\gamma - 1)\gamma^{\frac{(1 - \gamma)n}{2}}}{2(1 - \gamma)}.$$ 

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

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

and expand $n$ to get

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


Noting that \( x \) is minimized when \( 2^{(c-1) \log \left( \frac{1}{x} \right)} \) is maximized, and that this quantity is maximized when \( \log \left( \frac{1}{x} \right) \) is minimized (as \( c - 1 \) is negative), we get that \( \epsilon \)-optimality requires
\[
\epsilon \geq \frac{0.75 - 2^{1-c}}{4} y
\]
from \( \alpha < \frac{1}{8} \). Similarly, \( \alpha < \frac{1}{8} \) implies that \( \epsilon \)-optimality requires
\[
\epsilon \geq \frac{0.75 - 2^{-1}}{4} \left[ 1 - (2\gamma - 1) (\frac{1}{\gamma})^{\frac{1}{2(1-\gamma)}} \right].
\]
Note that \( 0.75 - 2^{1-c} \) is minimized when \( \gamma \) is small, so \( \gamma > c \) implies that \( \epsilon \)-optimality requires
\[
\epsilon \geq \frac{0.75 - 2^{-1}}{4} \left[ 1 - (2\gamma - 1) (\frac{1}{\gamma})^{\frac{1}{2(1-\gamma)}} \right] \geq \frac{1}{16} \left[ 1 - (2\gamma - 1) (\frac{1}{\gamma})^{\frac{1}{2(1-\gamma)}} \right].
\]
Conversely, \( 1 - (2\gamma - 1) (\frac{1}{\gamma})^{\frac{1}{2(1-\gamma)}} \) is minimized when \( \gamma \) is large, so as
\[
\lim_{\gamma \to 1} (2\gamma - 1) (\frac{1}{\gamma})^{\frac{1}{2(1-\gamma)}} = e^{c-1}
\]
we get that \( \epsilon \)-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. \( \square \)

### A.3. Omitted Proofs for Section 4

**Proof of Lemma 8.** We first show that either

- there exists an exploitation policy \( \pi \) in \( M_T \) such that
  \[
  \frac{1}{T} \max_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) - \frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) \leq \beta
  \]
  where the random variables \( \pi^t(s) \) and \( \tilde{\pi}^t(s) \) denote the states reached from \( s \) after following \( \pi \) and \( \tilde{\pi} \) for \( t \) steps, respectively,

- there exists an exploration policy \( \pi \) in \( M_T \) such that the probability that a walk of \( 2T \) steps from \( s \) following \( \pi \) will terminate in \( s_0 \) exceeds \( \frac{\beta}{2} \).

Let \( \pi \) be a policy in \( M \) satisfying
\[
\frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) = \frac{1}{T} \max_{\pi \in \Pi} \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) := \hat{V}.
\]
For any state \( s' \), let \( p(s') \) denote all the paths of length \( T \) in \( M \) that start in \( s' \), and \( q(s') \) 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 \( \Gamma \) 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 \( \Gamma \). Suppose
\[
\frac{1}{T} \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s)) < \hat{V} - \beta.
\]
Otherwise, \( \pi \) already witnesses the claim. We show that a walk of \( 2T \) steps from \( s \) following \( \pi \) will terminate in \( s_0 \) with probability of at least \( \frac{\beta}{2} \). First,
\[
\mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) = \mathbb{E} \sum_{t=1}^{T} \sum_{p(\pi^t(s))} \mathbb{P}[p(\pi^t(s))] V_M(p(\pi^t(s)))
\]
\[
= \mathbb{E} \sum_{t=1}^{T} \sum_{q(\pi^t(s))} \mathbb{P}[q(\pi^t(s))] V_M(q(\pi^t(s)))
\]
\[
+ \mathbb{E} \sum_{t=1}^{T} \sum_{r(\pi^t(s))} \mathbb{P}[r(\pi^t(s))] V_M(r(\pi^t(s)))
\]
since \( p(\pi^t(s)) = q(\pi^t(s)) \cup r(\pi^t(s)) \), which is a disjoint union. Next,
\[
\mathbb{E} \sum_{t=1}^{T} \sum_{q(\pi^t(s))} \mathbb{P}[q(\pi^t(s))] V_M(q(\pi^t(s)))
\]
\[
= \mathbb{E} \sum_{t=1}^{T} \sum_{q(\pi^t(s))} \mathbb{P}[q(\pi^t(s)) | M_T] V_M(q(\pi^t(s)))
\]
\[
\leq \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T),
\]
where the equality is due to Definition 9 and the definition of \( q \), and the inequality follows because \( V_M^\pi (\pi^t(s), T) \) is the sum over all the \( T \)-paths in \( M_T \), not just those that avoid the absorbing state \( s_0 \). Therefore by our original assumption on \( \pi \),
\[
\mathbb{E} \sum_{t=1}^{T} \sum_{q(\pi^t(s))} \mathbb{P}[q(\pi^t(s))] V_M(q(\pi^t(s)))
\]
\[
\leq \mathbb{E} \sum_{t=1}^{T} V_M^\pi (\pi^t(s), T) < T \hat{V} - T \beta.
\]
This implies

\[ E \sum_{t=1}^{T} \sum_{r(\pi^t(s))} P[r(\pi^t(s))] V_M(r(\pi^t(s))) \]

\[ = E \sum_{t=1}^{T} V_{\hat{M}}(\pi^t(s), T) \]

\[ - E \sum_{t=1}^{T} \sum_{q(\pi^t(s))} P[q(\pi^t(s))] V_M(q(\pi^t(s))) \]

\[ = TV - E \sum_{t=1}^{T} \sum_{q(\pi^t(s))} P[q(\pi^t(s))] V_M(q(\pi^t(s))) \geq T\beta, \]

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

\[ E \sum_{t=1}^{T} \sum_{r(\pi^t(s))} P[r(\pi^t(s))] V_M(r(\pi^t(s))) \]

\[ \leq T E \sum_{t=1}^{T} \sum_{r(\pi^t(s))} P[r(\pi^t(s))], \]

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

\[ T\beta \leq T^2 P_{2T}^\pi \]

and rearranging yields \( P_{2T}^\pi \geq \frac{\beta}{T} \) as desired.

Next, note the exploration policy (if it exists) can be derived by computing the optimal policy in \( \tilde{M}_T \). Moreover, the exploration policy (if it exists) in the exploitation MDP \( M_T \) can indeed be derived by computing the optimal policy in the exploration MDP \( M_T \) as observed by (Kearns and Singh, 2002). Finally, by Observation 5, any optimal policy in \( M_T^* (M_T^* \setminus \{\}) \) is an optimal policy in \( \tilde{M}_T (\tilde{M}_T \setminus \{\}) \).

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 \( \tilde{M} \) be two MDPs with the same set of states and actions. We say \( \tilde{M} \) is a \( \beta \)-approximation of \( M \) if

- For any state \( s \),
  \[ \tilde{R}_M(s) - \beta \leq \tilde{R}_\tilde{M}(s) \leq \tilde{R}_M(s) + \beta. \]

- For any states \( s \) and \( s' \) and action \( a \),
  \[ P_M(s, a, s') - \beta \leq P_\tilde{M}(s, a, s') \leq P_M(s, a, s') + \beta. \]

**Lemma 12** (Lemma 5, Kearns and Singh (2002)). Let \( M \) be an MDP and \( \Gamma \) 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 \( \bar{R}(s) \) denote the empirical estimates of the average reward obtained from visits to \( s \). Then with probability at least \( 1 - \delta \),

\[ |\hat{P}_M(s, a, s') - P_M(s, a, s')| < O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right), \]

and

\[ |\bar{R}_M(s) - \bar{R}_\tilde{M}(s)| < O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right). \]

Lemma 12 shows that \( \tilde{M}_T \) and \( \tilde{M}_{T|\Gamma} \) are \( O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right) \)-approximation MDPs for \( M_T \) and \( M_T \), respectively.

**Lemma 13** (Lemma 4, Kearns and Singh (2002)). Let \( M \) be an MDP and \( \tilde{M} \) its \( O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right) \)-approximation. Then for any policy \( \pi \in \Pi \) and any state \( s \) and action \( a \),

\[ V_\pi^\pi(s) - \min\{\epsilon, \alpha\} \leq V_\pi^\pi(s) \leq V_M^\pi(s) + \min\{\epsilon, \alpha\}, \]

and

\[ Q_\pi^\pi(s, a) - \min\{\epsilon, \alpha\} \leq Q_M^\pi(s, a) \leq Q_M^\pi(s, a) + \min\{\epsilon, \alpha\}. \]

**Proof of Lemma 10.** By Definition 7 and Lemma 12, \( \tilde{M}_T \) is a \( O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right) \)-approximation of \( M_T \). 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 \textbf{Fair-E}^3. To do so, we break down the probability of failure of \textbf{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 \( M_T \) to not be a \( O\left(\frac{\min\{\epsilon, \alpha\}^2}{n^2 H_T^3}\right) \)-approximation of \( M_T \).
2. At some known state the algorithm has a poor approximation of the \( Q_\pi^\pi \) values for one of the actions.
3. Following the exploration policy for \( 2T_{s}^* \) steps fails to yield enough visits to unknown states.
4. At some known state, the approximation value of that state in \( M_T \) is not an accurate estimate for the value of the state in \( M_T \).

We allocate \( \frac{\delta}{4} \) of our total probability of failure to each of these sources:
1. Set $\delta' = \frac{4}{4n}$ in Lemma 10.
2. Set $\delta' = \frac{2}{4nk}$ in Theorem 7.
3. By Lemma 8, each attempted exploration is a Bernoulli trial with probability of success of at least $\frac{n}{2T^*}$. 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 $H_T^*$. Therefore, the probability of taking fewer than $nm_Q$ trajectories of length $H_T^*$ would be bounded by $\frac{\delta}{4}$ if the number of $2T^*_M$ steps explorations is at least

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

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

### A.4. Relaxing Assumption 2

Throughout Sections 4.3 and 4.4 we assumed that $T^*_M$, the $\epsilon$-mixing time of the optimal policy $\pi^*$, was known (see Assumption 2). Although Fair-$E^3$ uses the knowledge of $T^*_M$ 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 $T^*_M$ and for any input $T^*_M$ runs in time $\text{poly}(T^*_M)$. Thus if $T^*_M$ is unknown, we can simply run Fair-$E^3$ for $T^*_M = 1, 2, \ldots$ sequentially and the running time and sample complexity will still be $\text{poly}(T^*_M)$. Similar to the analysis of Fair-$E^3$ when $T^*_M$ 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 $T^*_M$.

### B. Observations on Optimality and Fairness

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

**Proof.** In time $t$, let state $s_t$ denote the state from which $\pi$ chooses an action. Let $a^* = \arg \max_a Q^*_M(s_t, a)$ 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 $\pi^*$ be an optimal policy in MDP $M$. Then $\pi^*$ is approximate-action fair.

**Proof.** Assume that $\pi^*$ is not approximate-action fair. Given state $s$, the action that $\pi^*$ takes from $s$ is uniquely determined since $\pi^*$ is deterministic we may denote it by $a^*$. Then there exists a time step in which $\pi^*$ 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 $\pi^*$.

Observations 1 and 2 state that policies with optimal performance are fair; we now state that playing an action uniformly at random (regardless of the history) is fair.

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

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

**Observation 4.** Let $M$ be an MDP and $M^\alpha$ the $\alpha$-restricted MDP of $M$. Let $\pi$ be a policy in $M^\alpha$. Then $\pi$ is $\alpha$-action fair.

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

**Observation 5.** Let $M$ be an MDP and $M^\alpha$ the $\alpha$-restricted MDP of $M$. Let $\pi^*$ be an optimal policy in $M^\alpha$. Then $\pi^*$ is also optimal in $M$.

**Proof.** If $\pi^*$ 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 $M^\alpha$ by Definition 5.

### C. Omitted Details of Fair-$E^3$

We first formally define the exploitation MDP $M_E$ and the exploration MDP $M_{[\mu]}_E$:

$$[\mu]_E = [\mu]_E / \pi_E = [\mu]_E / \pi_E$$
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 $M_\Gamma = (\mathcal{S}_{M_\Gamma}, \mathcal{A}_{M_\Gamma}, P_{M_\Gamma}, R_{M_\Gamma}, T, \gamma)$ on $\Gamma$ where

- $S_{M_\Gamma} = \Gamma \cup \{s_0\}$.
- For any state $s \in \Gamma$, $\hat{R}_{M_\Gamma}(s) = \bar{R}_M(s)$, rewards in $M_\Gamma$ 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 $\Gamma$ are preserved in $M_\Gamma$.
- For any state $s_1 \in \Gamma$ and any action $a$, $P_{M_\Gamma}(s_1, a, s_0) = \Sigma_{s_2 \not\in \Gamma} P_M(s_1, a, s_2)$. Therefore, all the transitions between a state in $\Gamma$ and states not in $\Gamma$ are directed to $s_0$ in $M_\Gamma$.

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

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

Let $M$ be an MDP and $\Gamma$ 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 $\hat{R}_{M_\Gamma}(s)$ denote the empirical estimates of the average reward obtained from visits to $s$. Then $M_\Gamma$ is identical to $M_\Gamma$ except that:

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

Also $M_{\Gamma \setminus \Gamma}$ is identical to $M_{\Gamma \setminus \Gamma}$ except that:

- for any $s, s' \in \Gamma$ and action $a \in A$, $P_{M_{\Gamma \setminus \Gamma}}(s, a, s') = P_M(s, a, s')$. 