1. 1. first and second choices
   2. True
   3. False
   4. False
   5. second choice
   6. a, b, c, d

2. (a) The regret of the exploitation actions will decrease as the estimates gradually become better through more observations. However, since the exploration action explores arms uniformly at random, and epsilon is fixed, the probability of exploring bad arms doesn't decrease over time and regret will continue to be incurred.

   (b) By adding a confidence radius to the estimated mean rewards of each arm which depends on the horizon T and decreases with the number of times each arm has been played. Hence, when an arm is played, it was either selected because it has:

   - High confidence radius (exploration, we are uncertain about the estimated mean of this arm)
   - High estimated mean reward (exploitation, we are reasonably certain that this arm will yield high rewards)

3. (a) We first of need to calculate the expected reward from each state $E^\pi[r \mid s]$ (see Table 3) and the transition probabilities (see Table 2) We finally need to calculate

State	$E^\pi[r \mid s]$
0	$0.25 * 1 + 0.75*1 = 1$
1	$0.25 * 0 + 0.75*0 = 0$
2	$0.25 * 5 + 0.75*0 = 1.25$

Table 1: Expected reward for each state (when following the policy)

State	$P^\pi[s' = 0 \mid s]$	$P^\pi[s' = 1 \mid s]$	$P^\pi[s' = 2 \mid s]$
0	0.75	0.25	0
1	0.75	0	0.25
2	0.75	0	0.25

Table 2: Probability of transitioning to each next state from each state(when following the policy)

$$E^\pi[r_1 + r_2 \mid s_0 = s] = E^\pi[r \mid s] + \sum_{s_1=0}^{2} P(s_1 \mid s) E^\pi[r \mid s_1]$$

.

State	$E^\pi[r_1 + r_2 \mid s_0 = s]$
0	$1 + 0.75*1 + 0.25*0 = 1.75 = 7/4$
1	$0 + 0.75 * 1 + 0.25*1.25 = 1.0625 = 17/16$
2	$1.25 + 0.75 * 1 + 0.25*1.25 = 2.3125 = 37/16$

Table 3: Expected two-step reward for each state (when following the policy)

   (b) See slides/book for value iteration.

4. The issue arises since the policies do not have the same distributions of states (this is different from the general bandit case where the mixture of policies gives the same mixture of expected rewards). In the extreme case some states are never visited in either π^a or π^b can be visited in π^c. Two examples can be seen in Figure 1 where the reward is zero everywhere except one state which is only visited by π^c.

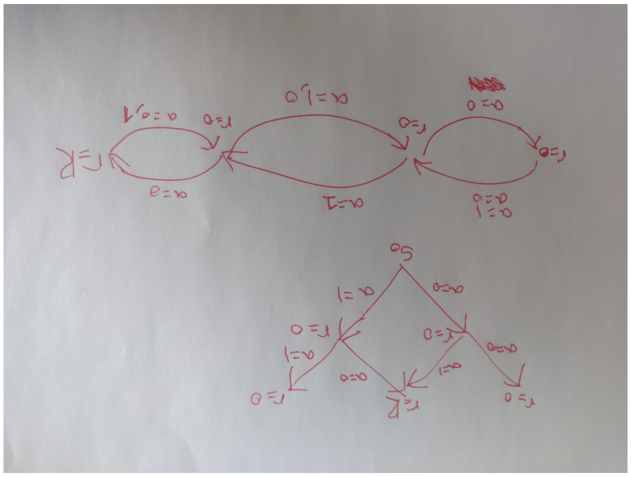

Figure 1: Caption

5. (a) First term: Expected difference of the upper confidence bound and lower confidence bound of the played arm at round t. Doesn't contain the mean of the played arm and is double the confidence radius, so if the confidence radius is sub-linear in T, so is this term.

   Second term: Expected value of how much the lower confidence bound exceeds the mean of the played arm at round t.

   Third term: How much the mean of the best arm exceeds the upper confidence bound in expectation.

   The last two terms correspond to the bad event in the analysis of the UCB algorithm, in the sense that if the probability of $\mu(a) \notin [\mathrm{LCB}_t(a), \mathrm{UCB}_t(a)]$ is low, these terms will be as well.

   (b) The third term can be bounded by the inequality in the following way:

$$
\begin{aligned}
\mathbb{E}\left[[\mu(a^*) - \mathrm{UCB}_t(a^*)]^+\right] \leq \quad & \mathbb{E}\left[\sum_{a \in \mathcal{A}} [\mu(a) - \mathrm{UCB}_t(a)]^+\right] \\
= \quad & \sum_{a \in \mathcal{A}} \mathbb{E}\left[[\mu(a) - \mathrm{UCB}_t(a)]^+\right] \\
= \quad & \sum_{a \in \mathcal{A}} \Bigg(\mathbb{E}\left[[\mu(a) - \mathrm{UCB}_t(a)]^+ \,\middle|\, \mu(a) \geq \mathrm{UCB}_t(a)\right] \cdot \Pr\{\mu(a) \geq \mathrm{UCB}_t(a)\} + \\
& \mathbb{E}\left[[\mu(a) - \mathrm{UCB}_t(a)]^+ \,\middle|\, \mu(a) < \mathrm{UCB}_t(a)\right] \cdot \Pr\{\mu(a) < \mathrm{UCB}_t(a)\} \Bigg) \\
= \quad & \sum_{a \in \mathcal{A}} \Bigg(\mathbb{E}\left[[\mu(a) - \mathrm{UCB}_t(a)]^+ \,\middle|\, \mu(a) \geq \mathrm{UCB}_t(a)\right] \cdot \Pr\{\mu(a) \geq \mathrm{UCB}_t(a)\} + \\
& 0 \cdot \Pr\{\mu(a) < \mathrm{UCB}_t(a)\} \Bigg) \\
= \quad & \sum_{a \in \mathcal{A}} \mathbb{E}\left[[\mu(a) - \mathrm{UCB}_t(a)]^+ \,\middle|\, \mu(a) \geq \mathrm{UCB}_t(a)\right] \cdot \Pr\{\mu(a) \geq \mathrm{UCB}_t(a)\} \\
= \quad & \sum_{a \in \mathcal{A}} \mathbb{E}\left[\mu(a) - \mathrm{UCB}_t(a) \,\middle|\, \mu(a) \geq \mathrm{UCB}_t(a)\right] \cdot \Pr\{\mu(a) \geq \mathrm{UCB}_t(a)\} \\
\leq \quad & \sum_{a \in \mathcal{A}} \frac{2}{TK} \\
= \quad & \frac{2}{T}
\end{aligned}
$$