Tutorial problems - Bandits. AV499-AVD871

1) Consider a 3-armed Bandit problem set up as follows:

Actions : $\{a_1, a_2, a_3\}$.

Reward distributions $R(s, a_1)$ is $N(1, 1)$.
$R(s, a_2)$ is $N(2, 2)$.
$R(s, a_3)$ is $N(3, 3)$.

Consider a randomized policy which picks the action A_t in slot t according to the distribution $\{0.2, 0.3, 0.5\}$ on $\{a_1, a_2, a_3\}$. (i.e. $Pr\{A_t = a_1\} = 0.2$ etc). Suppose A_t is picked independently for every t. Find out the expected regret for 10 slots.

2) Please review (and redo) the tutorial problem on Bandits discussed in class with the modification that the initial estimate $Q_0(a)$ was 5, $\forall a$.

3) Contextual Bandit problem:

Consider the following 2-state 2 armed contextual bandit. The states are $\{1, 2\}$ and the actions are $\{a_1, a_2\}$. The states are picked according to the uniform distribution on $\{1, 2\}$. Assume that the estimates $Q_t(s, a)$ are initialized for $t = 0$ as follows.

s	a	$Q_0(s,a)$
1	a_1	5
2	a_1	6
1	a_2	7
2	a_2	3

Assume that a greedy policy (fully greedy) is used for 2 steps. The reward values obtained are 4 and 10. ~~Show~~ Find out $Q_1(s, a)$ and $Q_2(s, a)$.