Question 1

(a) No, the states in blackjack are on the basis of player's current sum, the dealer's one showing card and whether or not holding a usable ace, So the same state never occur twice in the same episode. So there is no difference between first-visit and every-visit MC methods.

(b) 1) We can use MDP:

$$\begin{cases} V(E) = \frac{1}{2}\times 1 + \frac{1}{2}V(D) \\ V(D) = \frac{1}{2}V(C) + \frac{1}{2}V(E) \\ V(C) = \frac{1}{2}V(B) + \frac{1}{2}V(D) \\ V(B) = \frac{1}{2}V(A) + \frac{1}{2}V(C) \\ V(A) = \frac{1}{2}\times 0 + \frac{1}{2}V(B) \end{cases} \Rightarrow \begin{cases} V(E) = \frac{1}{2} + \frac{1}{2}V(D) \\ V(E) = 5V(A) \\ V(D) = 4V(A) \\ V(C) = 3V(A) \\ V(B) = 2V(A) \end{cases} \Rightarrow \begin{cases} V(A) = \frac{1}{6} \\ V(B) = \frac{2}{6} \\ V(C) = \frac{3}{6} \\ V(D) = \frac{4}{6} \\ V(E) = \frac{5}{6} \end{cases}$$

2) We can use TD(0) to update the value fuction of each state. By initializing V to 0 and setting $\alpha = 0.1$; After enough episodes, we can get V_* for each state.

I think for computing the truth value, using MDP can get V_* without generating episodes, which are also the precise values, as we know $p(s', r|s, a)$ for each state. So I guess we actually use this method.

(c) 1. When agent can do 8 possible actions, the episodes it can make within 8000 steps is twice the number of 4 actions.

2. The agent with 9 possible actions is worse than the agent with 8, but it still do better than 4 actions.

$\alpha = 0.5$, $\varepsilon = 0.1$, undiscounted.

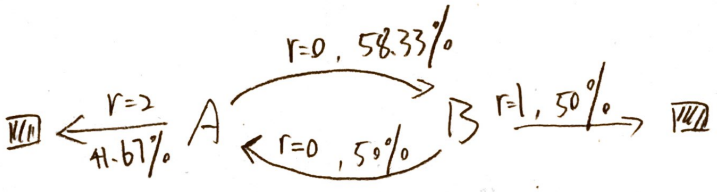
(d) Q-learning is off-policy because it updates the Q-values by using the Q-value of next state s' and take $\arg\max_a Q(s',a)$ regardless to the actual action chosen for next state s', which means the target policy is different from behavior policy.

Question 2

1. $$V(A) = \frac{1+2+2+1+2+1+2+1+2}{9} = \frac{14}{9}$$

$$V(B) = \frac{1+2+1+2+1+2+1+1+1+1+2}{11} = \frac{15}{11}$$

2



$A \xrightarrow{r=0} B \Rightarrow 7 \Rightarrow \frac{7}{7+5} = \frac{7}{12}$ $\quad B \xrightarrow{r=0} A \Rightarrow 7 \Rightarrow \frac{7}{14} = \frac{1}{2}$

$A \xrightarrow{r=2}$ ▨ $\Rightarrow 5 \Rightarrow \frac{5}{7+5} = \frac{5}{12}$ $\quad B \xrightarrow{r=1}$ ▨ $\Rightarrow 7 \Rightarrow \frac{7}{14} = \frac{1}{2}$

3

$$V(A) = 2 \times \frac{5}{12} + \frac{7}{12}(0 + V(B))$$

$$V(B) = 1 \times \frac{1}{2} + \frac{1}{2}(0 + V(A))$$

$$\Rightarrow \begin{cases} V(A) = \frac{27}{17} \\ V(B) = \frac{22}{17} \end{cases}$$