Question 1

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

(b) 1) We can use MDP:

$$V(E) = \frac{1}{2} \times 1 + \frac{1}{2} V(D)$$

$$V(D) = \frac{1}{2} V(c) + \frac{1}{2} V(E)$$

$$V(C) = \frac{1}{2} V(B) + \frac{1}{2} V(D)$$

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

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

 $V(E) = \frac{1}{2} + \frac{1}{2}V(D)$   $V(A) = \frac{1}{6}$   $V(E) = \frac{1}{2} + \frac{1}{2}V(D)$   $V(B) = \frac{1}{6}$   $V(B) = \frac{1}{2}V(A)$   $V(C) = \frac{1}{6}$   $V(C) = \frac{1}{6}$ 

2) We can use TD(0) to update the value fuction of each state. By initializing V to D and setting d = 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 episedes, which are also the precise values, as we know p(s', r/s, a) for each state. So I guess we adually 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.

Q = 0.5, E = 0.1, undiscounted.

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

Question 2

1. 
$$V(A) = \frac{1+2+2+1+2+1+2}{9} = \frac{19}{9}$$
  
 $V(B) = \frac{1+2+1+2+1+2+1+1+1+1+2}{11} = \frac{15}{11}$ 

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

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

$$= 7 \left( V(A) = \frac{27}{17} + \frac{2}{17} + \frac{2}{17$$

|     |  | , |
|-----|--|---|
|     |  |   |
|     |  |   |
| •   |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  | • |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  | 0 |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  | a |
|     |  |   |
|     |  |   |
|     | 원인경기원 중에 아기를 보고 있다고 그리다 그 모든 모든 그리는                        |   |
|     |  |   |
|     |  |   |
|     |  |   |
| 4.0 |  | • |
|     | 김경하 가게 된 보험에 맛있다. 그런 뭐 그들은 수에 되는 없이다. 그는 것 같               |   |
|     | 경찰과 경기 회사는 그 차를 되어 있었다. 그는 그리고 하시면 중심한 것이다.                |   |
|     |  |   |
|     | 지지 않는 이 중에 가는 모든 아님이 없이 되어 있다. 그리고 그리고 없는 그                |   |
|     |  | • |
|     | [[유리] [[ [ [ [ [ [ [ [ [ [ [ [ [ [ [ [ [ [                 |   |
|     | 그렇게 되는 하는 이 남이 들어 하는데 그 살이다. 그 중에 남은 그래 이 그릇은              |   |
|     | 요즘 발생이 휴대가 들어들으로 가능했다면서는 되다는 이번 경험 모양                      |   |
|     | 그는 살고 살면 무를 보는 이렇게 그렇게 하는 이 사람들이 하는데 하다.                   |   |
|     | 그 그렇게 되어 있는데 이렇게 되고 되었다. 양주의 이렇게 되고 있는 것은                  |   |
|     | 그렇게 하는 사람의 사람이 되고 이 시간이 되었다. 살아 살아 보다 다                    |   |
|     | 그 이번 회사 개발이 하나 있다면 하는 것이 되었다. 그리고 하는 그는 그 그는 그 그리고 하는 것이다. |   |
|     | 병원 항상을 시간하다 항목 맛있다면 하는 것이라면 하는 것이다.                        |   |
|     | 하고 있다. 그 이 나는 이 그리는 그 사이를 하셨다는 이 점이 되었다. 그                 |   |
|     | (3) 이번에 이번째 : "이 귀하나 있는데 아니는 아니는 이 것이다.                    |   |
|     |  | · |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |
|     |  |   |