1.

a) Consider the rules of Blackjack, along with the policy
(Stick only when having a sum of 20 or 21):

The last two rows in the rear represent the two states of player having a sum of 20 and 21. According to the policy, the action taken at these two states (always stick) will never lead to a bust, and also the chances of winning are also very high (close to 21), therefore the state-value function for these two states are also high.

The whole last row on the left represents the states 12 to 19. Since one deck of cards are mostly made up with face values of 10 (10, J, Q, K), according to the policy the agent always hit at states 12 to 19, there will be a very high chance of going bust, therefore the state-value function of states 12 to 19 drop off significantly.

The upper diagrams represents the player with an ace, and the lower diagrams represents the player without an ace. With an ace, the chances of either Blackjack or reaching state 20 is higher than without an ace. As previously explained, one deck of cards are mostly made up with face values of 10, in this case we even have extra four 9's as a usable card for sticking (not going bust), therefore the values in upper diagram is higher than lower diagram since there are higher chances of winning.

b) Method 1:
Directly applying Bellman's Equation
$v_{current} = (1/2)\, v_{left} + (1/2)\, v_{right}$

Method 2:
Generating many numbers of episodes and sample the return values, learn by using TD Prediction

In this case, method 2 is used. Assume a larger problem size, method 1 has a large computational cost, and unable to obtain the actual probability distribution. Using TD can directly compute the state on each time step, which yields better performance.

c) Programming

d) Since the episodes are generated under epsilon-greedy policy, and Q-Learning is a greedy algorithm, therefore Q-Learning is considered an off-policy control method. In addition, Q-Learning algorithm only updates the state, but cannot determine the action to be taken at the next time step.

Question 2. [24 points] (episodic example of TD and MC)

Suppose you observe the following 9 episodes generated by an unknown Markov reward process, where A and B are states and the numbers are rewards:

1. A,0,B,6	2. B,0,A,2,B,2	3. B,6
4. A,3	5. B,0,A,3	6. B,2
7. A,2,B,0,A,3	8. B,2	9. B,6

1. (8 pts) Give the values for states A and B that would be obtained by the batch first-visit Monte-Carlo method using this data set (assuming no discounting). You may express your answer using fractions. Briefly explain how you arrived at your answer.

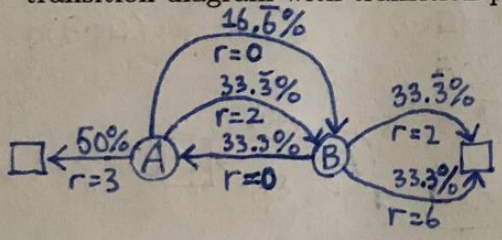
$v(A) = [6 + (2+2) + 3 + 3 + (2+3)] / 5 = 21/5 = 4.2$

(1st, 2nd, 4th, 5th, 7th; # of occurances)

$v(B) = [6 + (2+2) + 6 + 3 + 2 + 3 + 2 + 6] / 8 = 32/8 = 4$

(3rd, 6th, 8th, 9th)

2. (8 pts) If you were to form a maximum-likelihood model of a Markov reward process on the basis of these episodes (and these episodes alone), what would it be? (sketch its state-transition diagram with transition probabilities and expected rewards.)



$A \xrightarrow{3} \square$ 3 times $3/(3+1+2) = 50\%$

$A \xrightarrow{0} B$ 1 time $1/(3+1+2) = 16.\bar{6}\%$

$A \xrightarrow{2} B$ 2 times $2/(3+1+2) = 33.\bar{3}\%$

$B \xrightarrow{0} A$ 3 times $3/(3+3+3) = 33.\bar{3}\%$

$B \xrightarrow{2} \square$ 3 times same

$B \xrightarrow{6} \square$ 3 times same

3. (8 pts) Give the values for states A and B that would be obtained by the batch TD method. Briefly explain how you arrived at your answer. You may express your answer using fractions.

$v(A) = (\frac{1}{2})(3) + (\frac{1}{2})[(\frac{1}{3})(0) + (\frac{2}{3})(2) + v(B)] = \frac{3}{2} + \frac{2}{3} + \frac{v(B)}{2} = \frac{13 + 3v(B)}{6}$

$v(B) = (\frac{1}{3})(2) + (\frac{1}{3})(6) + (\frac{1}{3})[0 + v(A)] = \frac{8}{3} + \frac{v(A)}{3}$

$6v(A) = 13 + 3v(B)$

$3v(B) = 8 + v(A)$

$6v(A) = 21 + v(A)$

$v(A) = \frac{21}{5}$ $v(B) = \frac{61}{5} \cdot \frac{1}{3} = \frac{61}{15}$