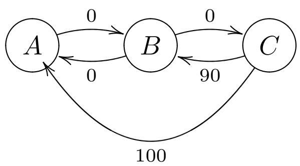
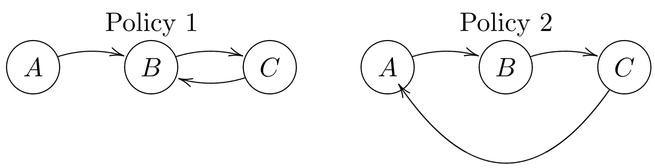
Homework-6 Solutions

Question 1



1

Guess two good policies for the above MDP. Draw them as a copy of the above diagram with arrows corresponding to the policy actions.



2

Guess the optimal policy for a discount rate of 0.0000001. Policy 2.

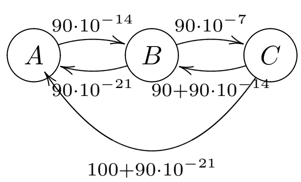
2.1

Here $\gamma = 0.0000001 = 1 \cdot 10^{-7}$. Compute V^* from your guess, then Q from V^* and then π^* from Q. Was your guess correct?

If the guess is Policy 1:

$$V^*(C) = 90 + 90\gamma^2 + 90\gamma^4 + \ldots = 90(1 + \gamma^2 + \gamma^4 + \ldots) = 90 \cdot \frac{1}{1-\gamma^2} \approx 90 + \epsilon \qquad \epsilon = 9 \cdot 10^{-13}$$
$$V^*(B) = \gamma V^*(C) \approx 90 \cdot 10^{-7}$$
$$V^*(A) = \gamma V^*(B) = 90 \cdot 10^{-14}$$

The corresponding Q:

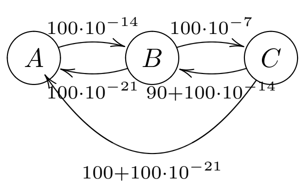


From this Q we get the optimal policy to be Policy 2. The edge from C to A has higher Q value than the edge from C to B. Therefore, our initial guess was wrong.

If the guess is Policy 2:

$$V^*(C) = 100 + 100\gamma^3 + 100\gamma^6 + \ldots == 100(1 + \gamma^3 + \gamma^6 + \ldots = 100 \cdot \frac{1}{1-\gamma^3} \approx 100 + \epsilon$$
$$V^*(B) = \gamma V^*(C) \approx 100 \cdot 10^{-7}$$
$$V^*(A) = \gamma V^*(B) = 100 \cdot 10^{-14}$$

The corresponding Q:



From this Q we get the optimal policy to be Policy 2. Therefore, our initial guess was right.

3

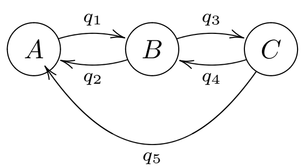
Guess the optimal policy for a discount rate of 0.9999999. Policy 1.

3.1

Here $\gamma = 0.9999999 = 1 - 1 \cdot 10^{-7}$. Compute V^* from your guess, then Q from V^* and then π^* from Q. Was your guess correct?

If the guess is Policy 1:

$$V^*(C) = 90 + 90\gamma^2 + 90\gamma^4 + \ldots = 90(1 + \gamma^2 + \gamma^4 + \ldots) = 90 \cdot \frac{1}{1-\gamma^2} \approx \frac{90}{2} \cdot 10^7$$
$$V^*(B) = \gamma V^*(C) \approx \frac{90}{2} \cdot (10^7 - 1)$$
$$V^*(A) = \gamma V^*(B) \approx \frac{90}{2} \cdot (10^7 - 2)$$

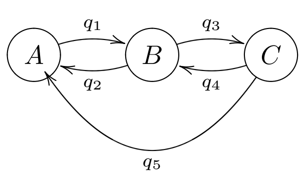
The corresponding Q: 

$$q_1 \approx \frac{90}{2} \cdot (10^7 - 2) \qquad q_2 \approx \frac{90}{2} \cdot (10^7 - 3) \qquad q_3 \approx \frac{90}{2} \cdot (10^7 - 1)$$
$$q_4 \approx 90 + \frac{90}{2} \cdot (10^7 - 2) \qquad q_5 \approx 100 + \frac{90}{2} \cdot (10^7 - 3)$$

Observe that $q_4 > q_5$. From this Q we get the optimal policy to be Policy 1. Therefore, our initial guess was right.

If the guess is Policy 2:

$$V^*(C) = 100 + 100\gamma^3 + 100\gamma^6 + \ldots = 100(1 + \gamma^3 + \gamma^6 + \ldots) = 100 \cdot \frac{1}{1-\gamma^3} \approx \frac{100}{3} \cdot 10^7$$
$$V^*(B) = \gamma V^*(C) \approx \frac{100}{3} \cdot (10^7 - 1)$$
$$V^*(A) = \gamma V^*(B) \approx \frac{100}{3} \cdot (10^7 - 2)$$

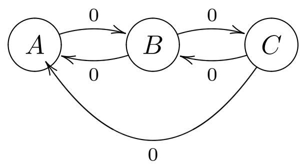
The corresponding Q: 

$$q_1 \approx \frac{100}{3} \cdot (10^7 - 2) \qquad q_2 \approx \frac{100}{3} \cdot (10^7 - 3) \qquad q_3 \approx \frac{100}{3} \cdot (10^7 - 1)$$
$$q_4 \approx 90 + \frac{100}{3} \cdot (10^7 - 2) \qquad q_5 \approx 100 + \frac{100}{3} \cdot (10^7 - 3)$$

Observe that $q_4 > q_5$.
From this Q we get the optimal policy to be Policy 1. Therefore, our initial guess was wrong.

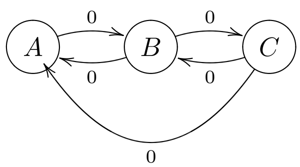
4

Run the Q learning algorithm for this problem with a discount rate of $1/2$. Start with $\hat{Q}(s,a) = 0$, which can be described by the following diagram:

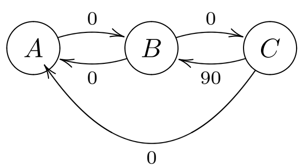


Produce the value of Q after the following pairs are considered one after the other:

1. (B,left).

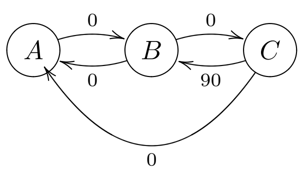


2. (C,left).

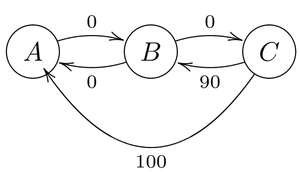


.

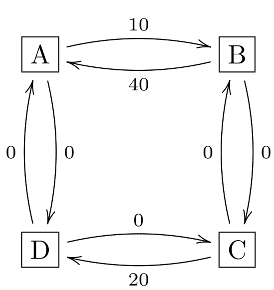
3. (A,right).



4. (C,down).



Question 2

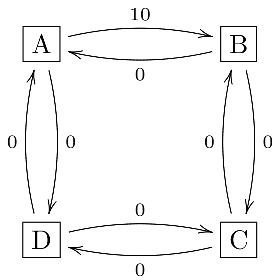


The following questions are related to the reinforcement online Q-learning algorithm applied to the above directed graph. The weights indicate rewards. Whenever needed use a discount rate of 0.5.

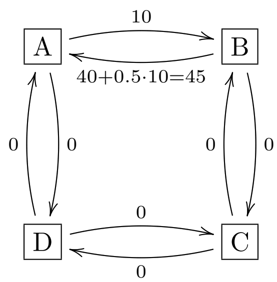
Part A

Run 4 iterations of the Q learning algorithm on this problem with a discount rate of 0.5. Start with $\hat{Q}(s,a) = 0$ for all states and actions, and compute the value of $\hat{Q}$ after the given actions are considered one after the other.

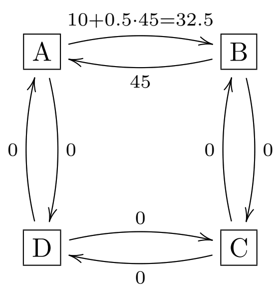
1. (A,right).



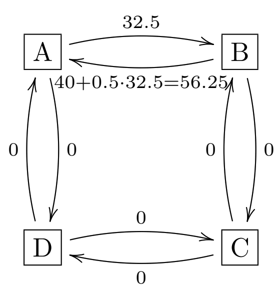
2. (B,left), after 1.



3. (A,right), after 2.



4. (B,left), after 3.



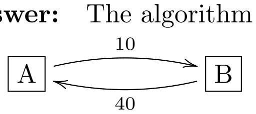
Part B

Starting with $\hat{Q}(s, a) = 0$ for all states and actions, the online Q-learning algorithm is applied repeatedly to the following actions infinitely many times. (No other actions are given to the algorithm.)

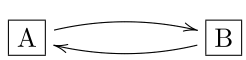
$$Q(B, \text{left}),\ Q(A, \text{right})$$

Do you expect the algorithm to converge to fixed values of $\hat{Q}$?

Answer: The algorithm is aware of the following subgraph:



The optimal policy π^* is clearly:



For the optimal policy the value of V^* is:

$$\begin{aligned} V^*(A) &= 10 + 1/2 \cdot 40 + (1/2)^2 \cdot 10 + (1/2)^3 \cdot 40 + \dots \\ &= 10 \cdot \frac{1}{1 - 1/4} + 40 \cdot 1/2 \cdot \frac{1}{1 - 1/4} = 40/3 + 80/3 = 40 \end{aligned}$$

$$V^*(B) = 40 + 1/2 \cdot V^*(A) = 40 + 20 = 60$$

Therefore, the value of Q^* is:
$$\begin{tikzcd} \boxed{A} \arrow[r, bend left, "40"] & \boxed{B} \arrow[l, bend left, "60"] \end{tikzcd}$$

With this input given repeatedly the Q values will:

converge to:
$$\begin{tikzcd} \boxed{A} \arrow[r, bend left, "40"] \arrow[d, bend left, "0"] & \boxed{B} \arrow[l, bend left, "60"] \arrow[d, bend left, "0"] \\ \boxed{D} \arrow[u, bend left, "0"] \arrow[r, bend left, "0"] & \boxed{C} \arrow[u, bend left, "0"] \arrow[l, bend left, "0"] \end{tikzcd}$$

Question 3

Consider the problem of clustering the $m = 4$ points below into $k = 2$ clusters.

1	(0,0)
2	(4,0)
3	(5,1)
4	(6,0)

Part 1

Suppose the points 3 and 4 are selected using the Lloyd technique for computing initial means.

1. What is the initial clustering computed by the k-means algorithm?

Answer:

$$u_1 = (5, 1), \qquad u_2 = (6, 0)$$

Squared distances from u_1, u_2:

$$\{26, 36\}, \quad \{2, 4\}, \quad \{0, 2\}, \quad \{2, 0\},$$

Clustering:

$$C(1) = 1, \quad C(2) = 1, \quad C(3) = 1, \quad C(4) = 2$$

2. What clustering is computed after the first iteration of the k-means algorithm?

Answer:

$$u_1 = (3, 1/3), \qquad u_2 = (6, 0)$$

Squared distances from u_1, u_2:

$$\{9 + 1/9, 36\}, \quad \{1 + 1/9, 4\}, \quad \{4 + 4/9, 2\}, \quad \{9 + 1/9, 0\},$$

Clustering:

$$C(1) = 1, \quad C(2) = 1, \quad C(3) = 2, \quad C(4) = 2$$

3. What clustering is computed after the second iteration of the k-means algorithm?

Answer:

$$u_1 = (2, 0), \qquad u_2 = (5.5, 0.5)$$

Squared distances from u_1, u_2:

$$\{4, 30.5\}, \quad \{4, 2.5\}, \quad \{10, 0.5\}, \quad \{16, 0.5\},$$

Clustering:

$$C(1) = 1, \quad C(2) = 2, \quad C(3) = 2, \quad C(4) = 2$$

4. What clustering is computed after the third iteration of the k-means algorithm?

Answer:

$$u_1 = (0, , 0), \qquad u_2 = (5, 1/3)$$

Squared distances from u_1, u_2:

$$\{0, 25 + 1/9\}, \quad \{16, 1 + 1/9\}, \quad \{26, 4/9\}, \quad \{36, 1 + 4/9\},$$

Clustering:

$$C(1) = 1, \quad C(2) = 2, \quad C(3) = 2, \quad C(4) = 2$$

5. What clustering is computed after the fourth iteration of the k-means algorithm?

Answer: Same as above.

6. What clustering is computed after the fifth iteration of the k-means algorithm?

Answer: Same as above.

Question 4

Consider the problem of clustering the $m = 6$ points below into $k = 2$ clusters.

1	(1,1)
2	(1,2)
3	(1,0)
4	(4,1)
5	(4,2)
6	(4,0)

Part 1

Here we will be using k-means.

1. What clstering is obtained by k-means if the initial points selected by the Lloyd technique are points 3 and 5? What is the corresponding quantization error?

Answer:

$$C(1) = 1, \; C(2) = 1, \; C(3) = 1, \; C(4) = 2, \; C(5) = 2, \; C(6) = 2, \qquad E = 4$$

2. What clstering is obtained by k-means if the initial points selected by the Lloyd technique are points 4 and 6? What is the corresponding quantization error?

Answer:

$$C(1) = 1, \; C(2) = 1, \; C(3) = 2, \; C(4) = 1, \; C(5) = 1, \; C(6) = 2, \qquad E = 14.5$$

Part 2

Here we will be using k-means++. Suppose the first point selected (at random) by the algorithm is Point 1.

1. Complete the following table for the squared distances of all points from Point 1:

1	$d = 0$
2	$d = 1$
3	$d = 1$
4	$d = 9$
5	$d = 10$
6	$d = 10$

2. Compute the probability of selecting each one of the points as the second point.

$$z = 0 + 1 + 1 + 9 + 10 + 10 = 31$$

1	$p = 0/31$
2	$p = 1/31$
3	$p = 1/31$
4	$p = 9/31$
5	$p = 10/31$
6	$p = 10/31$

3. Suppose the algorithm selects the point with the largest probability. What is the clstering obtained by k-means++? What is the corresponding quantization error?

Answer:

$$C(1) = 1,\ C(2) = 1,\ C(3) = 1,\ C(4) = 2,\ C(5) = 2,\ C(6) = 2, \qquad E = 4$$

4. Will your answer change if the algorithm selects the point with second largest probability?

Answer: No.

5. Will your answer change if the algorithm selects the point with third largest probability?

Answer: No.

6. What is the probability that the algorithm selects one of the three points with the largest probability?

Answer:

$$29/31 = 0.935...$$