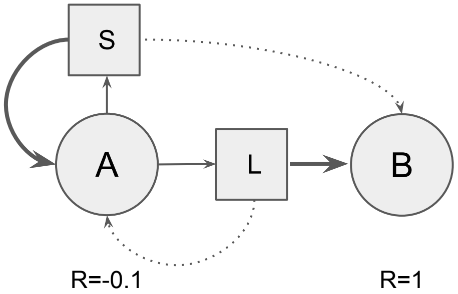
CS4246 / CS5446

Tutorial Week 9

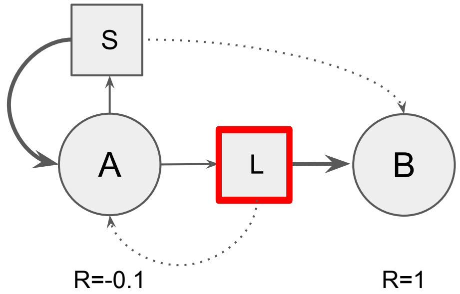
Muhammad **Rizki** Maulana
rizki@u.nus.edu

First



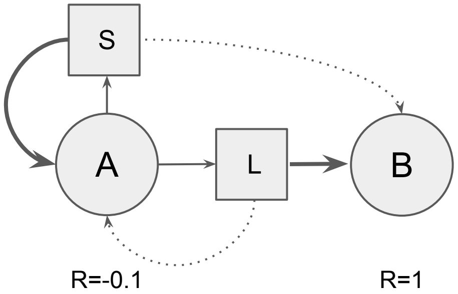
(a) Assume that actions L is more likely to succeed than not, and similarly action S is also more likely to succeed than not. What is the optimal policy π^*?

Question



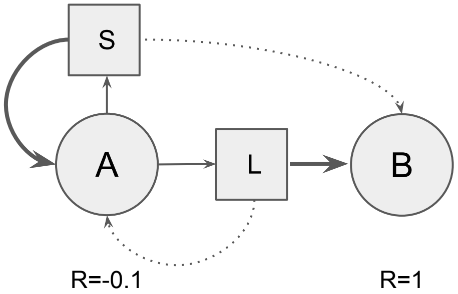
(a) Assume that actions L is more likely to succeed than not, and similarly action S is also more likely to succeed than not. What is the optimal policy π^*?

$\pi^*(A) = L.$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

Question



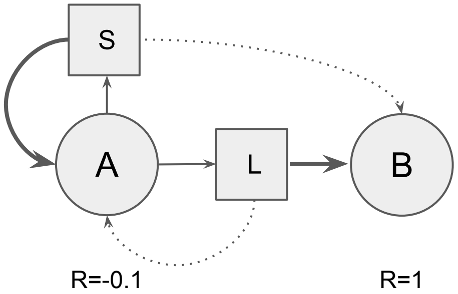
(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

AA
AA
AB

AA
AB

AB
AB

$T(A, L, A) = 3/7$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

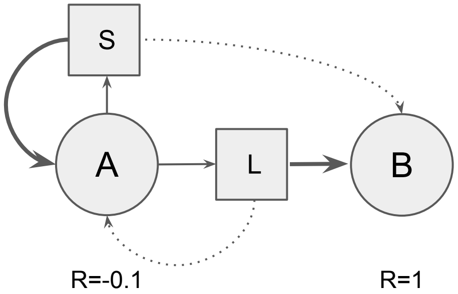
AA
AA
AB

AA
AB

AB
AB

$T(A, L, A) = 3/7$

$T(A, L, B) = 4/7.$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

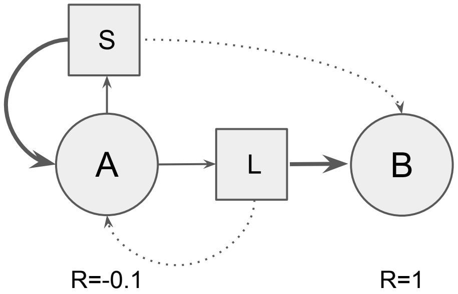
AA
AA
AB

AA
AB

AB
AB

$$T(A, L, A) = 3/7$$

$$T(A, L, B) = 4/7.$$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

AA
AA
AB

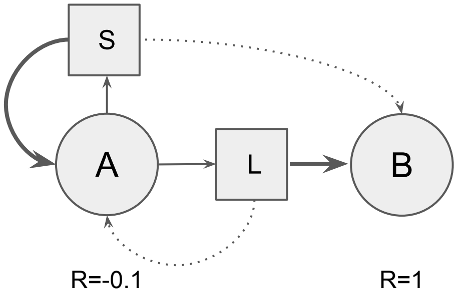
AA
AB

AB
AB

$$T(A, L, A) = 3/7$$

$$T(A, L, B) = 4/7.$$

$$U^{\pi^*}(A) = R(A) + \gamma \left(T(A, L, A)\, U^{\pi^*}(A) + T(A, L, B)\, U^{\pi^*}(B)\right)$$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

AA
AA
AB

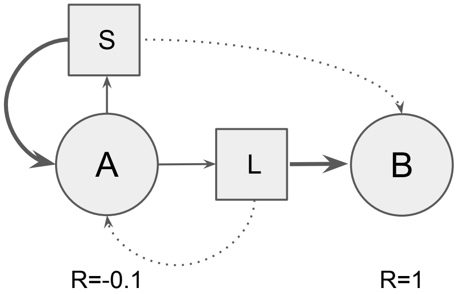
AA
AB

AB
AB

$$T(A, L, A) = 3/7$$

$$T(A, L, B) = 4/7.$$

$$U^{\pi^*}(A) = R(A) + \gamma \left(T(A, L, A)\, U^{\pi^*}(A) + T(A, L, B)\, U^{\pi^*}(B)\right)$$
$$U^{\pi^*}(A) = -0.1 + 0.5 \times (3/7 \times U^{\pi^*}(A) + 4/7 \times 1)$$



(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?

AA
AA
AB

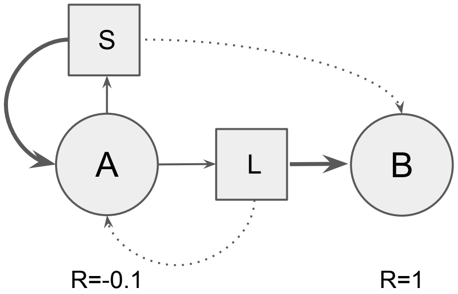
AA
AB

AB
AB

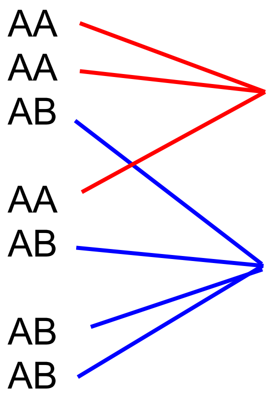
$$T(A, L, A) = 3/7$$

$$T(A, L, B) = 4/7.$$

$$U^{\pi^*}(A) = R(A) + \gamma\,(T(A, L, A)\,U^{\pi^*}(A) + T(A, L, B)\,U^{\pi^*}(B))$$
$$U^{\pi^*}(A) = -0.1 + 0.5 \times (3/7 \times U^{\pi^*}(A) + 4/7 \times 1)$$
$$11/14 \times U^{\pi^*}(A) = -0.1 + 4/14$$



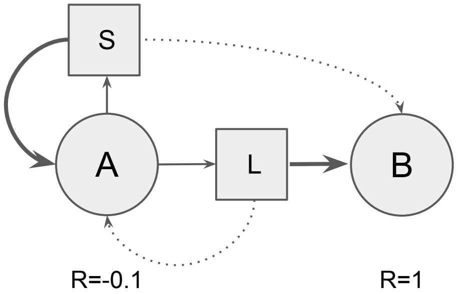
(b) Assume that the agent knows neither the transition function nor the utilities of the states. Assume that the agent, for some reason, happens to follow the optimal policy π^*. The rewards received at states A and B are the same as described above. In the process of executing this policy, the agent executes four trials and, in each trial, it stops after reaching state B. The following state sequences are recorded during the trials: $AAAB$, AAB, AB, AB. What is the estimate of $T(.,.,.)$? Using ADP, what is the estimate of $U^{\pi^*}(A)$, assuming a discount factor of $\gamma = 0.5$?



$T(A, L, A) = 3/7$

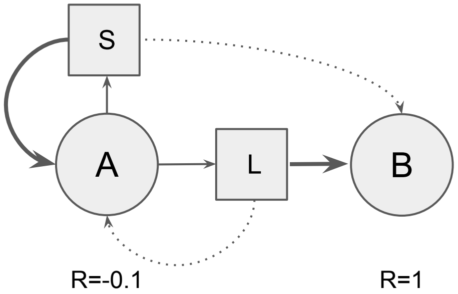
$T(A, L, B) = 4/7.$

$$U^{\pi^*}(A) = R(A) + \gamma\,(T(A, L, A)\,U^{\pi^*}(A) + T(A, L, B)\,U^{\pi^*}(B))$$
$$U^{\pi^*}(A) = -0.1 + 0.5 \times (3/7 \times U^{\pi^*}(A) + 4/7 \times 1)$$
$$11/14 \times U^{\pi^*}(A) = -0.1 + 4/14$$
$$U^{\pi^*}(A) = 26/110 = 0.2364.$$



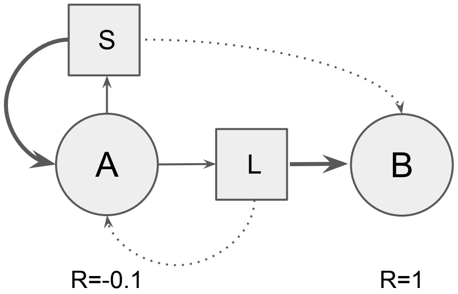
(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

Question



(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

$$U^{\pi^*}(A) \leftarrow \overbrace{U^{\pi^*}(A)}^{\text{Current estimate}} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \overbrace{U^{\pi^*}(A)})$$



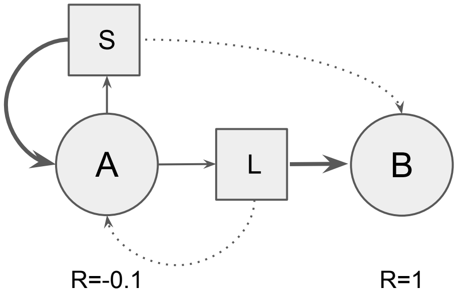
(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

Current estimate

$$U^{\pi^*}(A) \leftarrow \underbrace{U^{\pi^*}(A)} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \underbrace{U^{\pi^*}(A)})$$

AA

AB

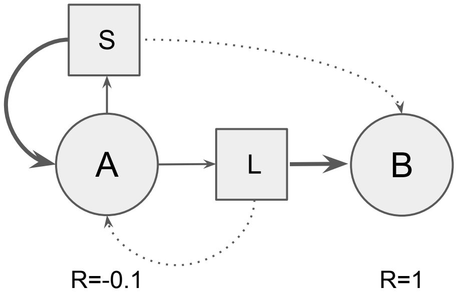


(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

$$U^{\pi^*}(A) \leftarrow \overbrace{U^{\pi^*}(A)}^{\text{Current estimate}} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \overbrace{U^{\pi^*}(A)})$$

AA $$U^{\pi^*}(A) \leftarrow U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(A) - U^{\pi^*}(A))$$

AB



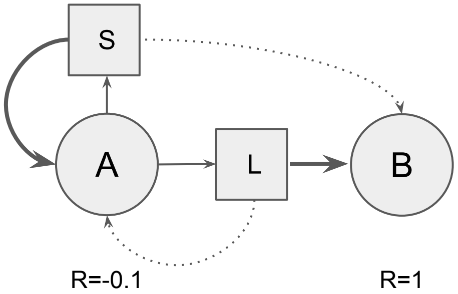
(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

$$U^{\pi^*}(A) \leftarrow \overbrace{U^{\pi^*}(A)}^{\text{Current estimate}} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \overbrace{U^{\pi^*}(A)})$$

AA

$$\begin{aligned} U^{\pi^*}(A) \leftarrow & \; U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(A) - U^{\pi^*}(A)) \\ = & \; -0.1 + 0.5 \times (-0.1 + 0.5 \times -0.1 - (-0.1)) = -0.125 \end{aligned}$$

AB



(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

Current estimate

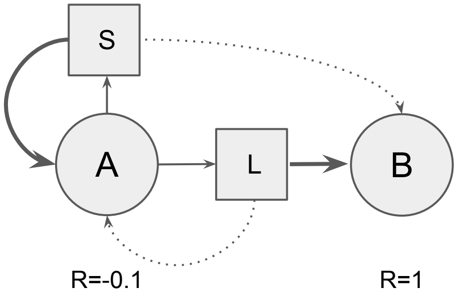
$$U^{\pi^*}(A) \leftarrow \overbrace{U^{\pi^*}(A)} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \overbrace{U^{\pi^*}(A)})$$

AA

$$\begin{aligned} U^{\pi^*}(A) &\leftarrow U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(A) - U^{\pi^*}(A)) \\ &= -0.1 + 0.5 \times (-0.1 + 0.5 \times -0.1 - (-0.1)) = -0.125 \end{aligned}$$

AB

$$U^{\pi^*}(A) \leftarrow U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(B) - U^{\pi^*}(A))$$



(c) Assume now that the agent is executing only one trial yielding the sequence of states AAB. Compute the estimate of the utility $U^{\pi^*}(A)$ using TD learning. Use discount $\gamma = 0.5$ and learning rate $\alpha = 0.5$. To know the starting values for U^{π^*}, refer to the TD learning algorithm in the lecture notes.

$$U^{\pi^*}(A) \leftarrow \overbrace{U^{\pi^*}(A)}^{\text{Current estimate}} + \alpha(\underbrace{R(A) + \gamma U^{\pi^*}(B)}_{\text{TD Target}} - \overbrace{U^{\pi^*}(A)}^{})$$

AA

$$\begin{aligned} U^{\pi^*}(A) &\leftarrow U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(A) - U^{\pi^*}(A)) \\ &= -0.1 + 0.5 \times (-0.1 + 0.5 \times -0.1 - (-0.1)) = -0.125 \end{aligned}$$

AB

$$\begin{aligned} U^{\pi^*}(A) &\leftarrow U^{\pi^*}(A) + \alpha(R(A) + \gamma U^{\pi^*}(B) - U^{\pi^*}(A)) \\ &= -0.125 + 0.5 \times (-0.1 + 0.5 \times 1 - (-0.125)) = 0.1375 \end{aligned}$$

Second

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

Q	s_1	s_2
a_1	0	0
a_2	0	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s', a') - Q(s,a))$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

Q	s_1	s_2
a_1	0	0
a_2	0	0

$$Q(s,a) \leftarrow Q(s,a) + \alpha\Big(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}}\Big)$$

(a) s_1, $R(s_1) = -10$, a_1, s_1

Question

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	$\boxed{0}$	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$\begin{aligned} Q(s_1, a_1) &\leftarrow 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ &= -5 \end{aligned}$$

Q	s_1	s_2
a_1	$\boxed{-5}$	0
a_2	0	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\overbrace{(\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a))}^{\text{Error}}$$

Q	s_1	s_2
a_1	$\boxed{0}$	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$Q(s_1,a_1) \leftarrow 0 + 0.5(-10 + 0.5\max(0,0) - 0)$$
$$= -5$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

Question

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$\begin{aligned} Q(s_1,a_1) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

$$\begin{aligned} Q(s_1,a_2) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	-5	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$\begin{aligned} Q(s_1,a_1) \leftarrow\ & 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ =\ & -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

$$\begin{aligned} Q(s_1,a_2) \leftarrow\ & 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ =\ & -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	-5	0

(c) s_2, $R(s_2) = 20$, a_1, s_1

Question

Q-Learning.
Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$\begin{aligned} Q(s_1,a_1) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

$$\begin{aligned} Q(s_1,a_2) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	-5	0

(c) s_2, $R(s_2) = 20$, a_1, s_1

$$\begin{aligned} Q(s_2,a_1) \leftarrow &\ 0 + 0.5(20 + 0.5\max(-5,-5) - 0) \\ = &\ 8.75 \end{aligned}$$

Q	s_1	s_2
a_1	-5	8.75
a_2	-5	0

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$Q(s_1,a_1) \leftarrow 0 + 0.5(-10 + 0.5\max(0,0) - 0) = -5$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

$$Q(s_1,a_2) \leftarrow 0 + 0.5(-10 + 0.5\max(0,0) - 0) = -5$$

Q	s_1	s_2
a_1	-5	0
a_2	-5	0

(c) s_2, $R(s_2) = 20$, a_1, s_1

$$Q(s_2,a_1) \leftarrow 0 + 0.5(20 + 0.5\max(-5,-5) - 0) = 8.75$$

Q	s_1	s_2
a_1	-5	8.75
a_2	-5	0

(d) s_1, $R(s_1) = -10$, a_2, s_2

Question

Q-Learning.

Consider a system with two states s_1, s_2 and two actions a_1, a_2. You perform actions and observe the rewards and state transitions listed below. Each step lists the current state, observed reward, action, and resulting next state as s_i, $R(s_i) = r$, a_k, and s_j, respectively. Perform Q-learning using a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$ for each step. The Q-value entries in the Q-table are initialized to zero.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\overbrace{\underbrace{R(s) + \gamma \max_{a'} Q(s',a')}_{\text{Target}} - Q(s,a)}^{\text{Error}})$$

Q	s_1	s_2
a_1	0	0
a_2	0	0

(a) s_1, $R(s_1) = -10$, a_1, s_1

$$\begin{aligned} Q(s_1,a_1) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	0	0

(b) s_1, $R(s_1) = -10$, a_2, s_2

$$\begin{aligned} Q(s_1,a_2) \leftarrow &\ 0 + 0.5(-10 + 0.5\max(0,0) - 0) \\ = &\ -5 \end{aligned}$$

Q	s_1	s_2
a_1	-5	0
a_2	-5	0

(c) s_2, $R(s_2) = 20$, a_1, s_1

$$\begin{aligned} Q(s_2,a_1) \leftarrow &\ 0 + 0.5(20 + 0.5\max(-5,-5) - 0) \\ = &\ 8.75 \end{aligned}$$

Q	s_1	s_2
a_1	-5	8.75
a_2	-5	0

(d) s_1, $R(s_1) = -10$, a_2, s_2

$$\begin{aligned} Q(s_1,a_2) \leftarrow &\ -5 + 0.5(-10 + 0.5\max(8.75,0) - (-5)) \\ = &\ -5.3125 \end{aligned}$$

Q	s_1	s_2
a_1	-5	8.75
a_2	-5.3125	0

Third

SARSA and Q-learning.

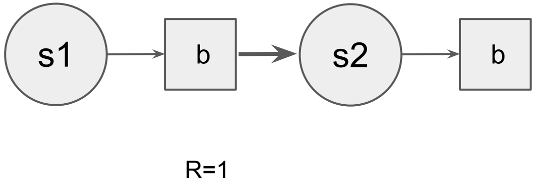
Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

SARSA and Q-learning.
Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

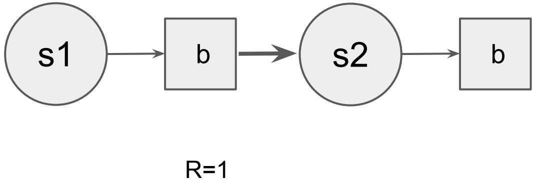
Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.



SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:



Q	s_1	s_2
a	2	4
b	2	2

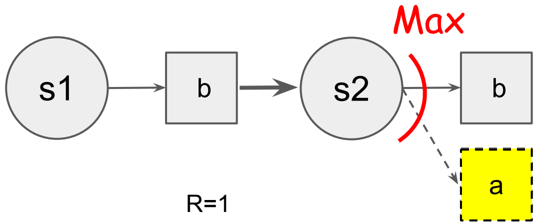
Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.

SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.



Q-Learning:

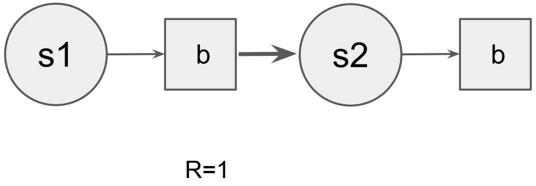
$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.



Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \quad Q(s',a') - Q(s,a))$$

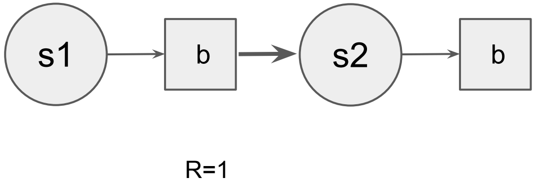
Question

SARSA and Q-learning.
Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	$\boxed{2}$	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.

$Q(s_1, b)$ is the affected entry.



Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \qquad Q(s',a') - Q(s,a))$$

SARSA and Q-learning.

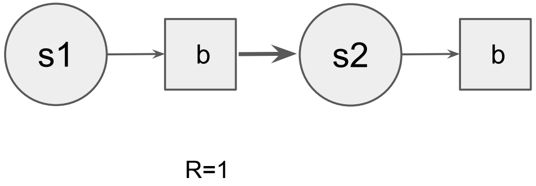
Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	$\boxed{2}$	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.

$Q(s_1, b)$ is the affected entry.

For SARSA,



Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

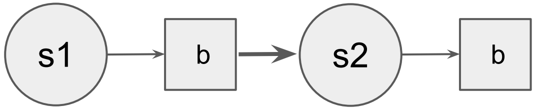
$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \qquad Q(s',a') - Q(s,a))$$

SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.



R=1

Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \quad Q(s',a') - Q(s,a))$$

$Q(s_1, b)$ is the affected entry.

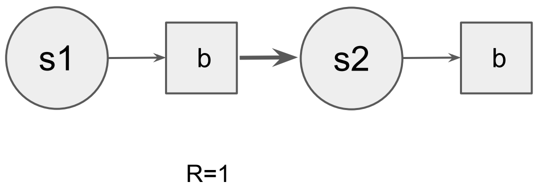
For SARSA,

$$\begin{aligned} Q(s_1,b) \leftarrow & \; Q(s_1,b) + \alpha(R(s_1) + \gamma \boxed{Q(s_2,b)} - Q(s_1,b)) \\ = & \; 2 + 0.2 \times (1 + 0.8 \times 2 - 2) = 2.12 \end{aligned}$$

SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	$\boxed{2}$	2



Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.

Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \quad Q(s',a') - Q(s,a))$$

$Q(s_1, b)$ is the affected entry.

For SARSA,

$$\begin{aligned} Q(s_1,b) \leftarrow & \quad Q(s_1,b) + \alpha(R(s_1) + \gamma Q(s_2,b) - Q(s_1,b)) \\ = & \quad 2 + 0.2 \times (1 + 0.8 \times 2 - 2) = 2.12 \end{aligned}$$

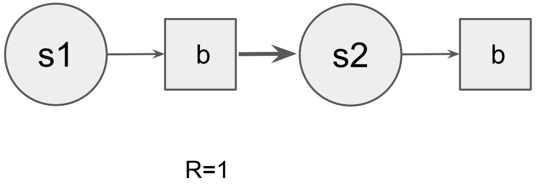
For Q-learning,

SARSA and Q-learning.

Consider using SARSA and Q-learning to learn a policy in an MDP with two states s_1 and s_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$, and that the current values of Q are:

Q	s_1	s_2
a	2	4
b	2	2

Suppose that, when we were in state s_1, we took action b, received reward 1 and moved to state s_2 and take action b there. Which item of the Q-table will change and what is the new value? Compute for both SARSA and Q-learning.



Q-Learning:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

SARSA:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \quad Q(s',a') - Q(s,a))$$

$Q(s_1, b)$ is the affected entry.

For SARSA,

$$\begin{aligned} Q(s_1, b) \leftarrow & \quad Q(s_1, b) + \alpha(R(s_1) + \gamma Q(s_2, b) - Q(s_1, b)) \\ = & \quad 2 + 0.2 \times (1 + 0.8 \times 2 - 2) = 2.12 \end{aligned}$$

For Q-learning,

$$\begin{aligned} Q(s_1, b) \leftarrow & \quad Q(s_1, b) + \alpha(R(s_1) + \gamma \boxed{\max_{u \in \{a,b\}} Q(s_2, u)} - Q(s_1, b)) \\ = & \quad 2 + 0.2 \times (1 + 0.8 \times 4 - 2) = 2.44 \end{aligned}$$

Question?

<EOF>