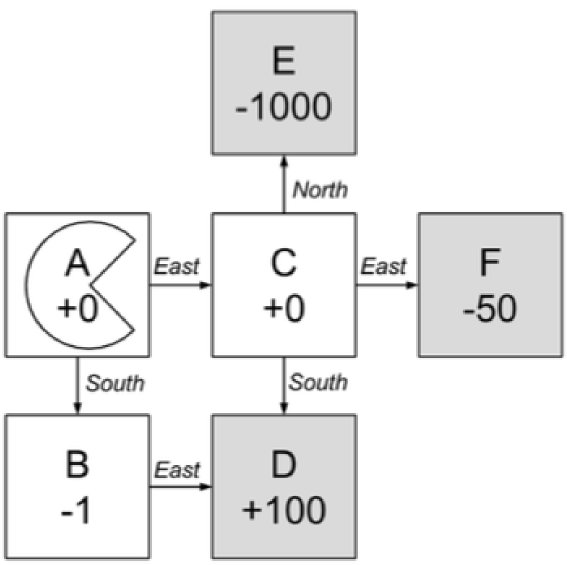
Q8: Indecisive Pacman

Q8: Indecisive Pacman

Simple MDP

Pacman is an agent in a deterministic MDP with states A, B, C, D, E, F. He can deterministically choose to follow any edge pointing out of the state he is currently in, corresponding to an action $North$, $East$, or $South$. He cannot stay in place. D, E, and F are terminal states. Let the discount factor be $\gamma = 1$. Pacman receives the reward value labeled underneath a state upon entering that state.



Part 1

0.0/3.0 points (graded)

Write the optimal values $V^*(s)$ for $s = A$ and $s = C$ and the optimal policy $\pi^*(s)$ for $s = A$.

$V^*(A) =$

100 **Answer:** 100

$V^*(C)$

100 **Answer:** 100

$\pi^*(A)$

East ▾ **Answer:** East

Explanation

The optimal plan without indecisiveness is to go East and then South, collecting no negative rewards.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Part 2

0.0/2.0 points (graded)

Pacman is typically rational, but now becomes indecisive if he enters state C. In state C, he finds the two best actions and randomly, with equal probability, chooses between the two. Let $\bar{V}(s)$ be the values under the policy where Pacman acts according to $\pi^*(s)$ for all $s \neq C$, and follows the indecisive policy when at state C. What are the values $\bar{V}(s)$ for $s = A$ and $s = C$?

$\bar{V}(A)$

Generating Speech Output

25

Answer: 25

$\bar{V}(C)$

25

Answer: 25

Explanation

Pacman is unaware of his indecisiveness at state C. So he will follow his policy π^* from (i) at state A and go East. When he has reached C, his two best actions are going South and going East. He will then receive the average of the value of taking those two actions, $(100 - 50)/2 = 25$, since he will take either one with equal probability.

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

Part 3

0.0/2.0 points (graded)

Now Pacman knows that he is going to be indecisive when at state C and decides to recompute the optimal policy at all other states, anticipating his indecisiveness at C. What is Pacman's new policy $\tilde{\pi}(s)$ and new value $\tilde{V}(s)$ for $s = A$?

$\tilde{\pi}(A)$

South

Answer: South

$\tilde{V}(A)$

99

Answer: 99

Explanation

Now Pacman knows about his indecisiveness as C, and so he can anticipate the low value (25) of being in C. He can therefore choose to go South and receive a reward of -1 and from there go East, to get a reward of 100, making for a value from A of 99.

Generating Speech Output

Submit

You have used 0 of 2 attempts

Answers are displayed within the problem

General Case - Indecisive everywhere

Pacman enters a new non-deterministic MDP and has become indecisive in all states of this MDP: at every time-step, instead of being able to pick a single action to execute, he always picks the two distinct best actions and then flips a fair coin to randomly decide which action to execution from the two actions he picked.

Let S be the state space of the MDP. Let $A(s)$ be the set of actions available to Pacman in state s. Assume for simplicity that there are always at least two actions available from each state ($|A(s)| \geq 2$).

This type of agent can be formalized by modifying the Bellman Equation for optimality. Let $\hat{V}(s)$ be the value of the indecisive policy. Precisely:

$$\hat{V}(s_0) = E[R(s_0, a_0, s_1) + \gamma R(s_1, a_1, s_2) + \gamma^2 R(s_2, a_2, s_3) + \ldots]$$

Let $\hat{Q}(s, a)$ be the expected utility of taking action a from state s and then following the indecisive policy after that step. We have that:

$$\hat{Q}(s, a) = \sum_{s' \in S} T(s, a, s')(R(s, a, s') + \gamma \hat{V}(s'))$$

Part 4

0.0/3.0 points (graded)

Which of the following options gives $\hat{V}$ in terms of $\hat{Q}$? When combined with the above formula for $\hat{Q}(s, a)$ in terms of $\hat{V}(s')$, the answer to this question forms the Bellman Equation for this policy.

$\hat{V}(s) =$

- $\max_{a \in A(s)} \hat{Q}(s, a)$

Generating Speech Output

- ( ) $\max_{a_1 \in A(s)} \max_{a_2 \in A(s),\, a_1 \neq a_2} \left(\hat{Q}(s, a_1) \cdot \hat{Q}(s, a_2)\right)$
- (•) $\max_{a_1 \in A(s)} \max_{a_2 \in A(s),\, a_1 \neq a_2} \frac{1}{2}\left(\hat{Q}(s, a_1) + \hat{Q}(s, a_2)\right)$ ✔
- ( ) $\max_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \left(\hat{Q}(s, a_1) \cdot \hat{Q}(s, a_2)\right)$
- ( ) $\sum_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \left(\hat{Q}(s, a_1) \cdot \hat{Q}(s, a_2)\right)$
- ( ) $\sum_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \frac{1}{2}\left(\hat{Q}(s, a_1) + \hat{Q}(s, a_2)\right)$
- ( ) $\max_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \frac{1}{2}\left(\hat{Q}(s, a_1) + \hat{Q}(s, a_2)\right)$
- ( ) $\frac{1}{|A(s)|(|A(s)|-1)} \sum_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \left(\hat{Q}(s, a_1) \cdot \hat{Q}(s, a_2)\right)$
- ( ) $\frac{1}{|A(s)|(|A(s)|-1)} \sum_{a_1 \in A(s)} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \frac{1}{2}\left(\hat{Q}(s, a_1) + \hat{Q}(s, a_2)\right)$
- ( ) $\max_{a_1 \in A(s)} \frac{1}{|A(s)|-1} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \frac{1}{2}\left(\hat{Q}(s, a_1) + \hat{Q}(s, a_2)\right)$
- ( ) $\max_{a_1 \in A(s)} \frac{1}{|A(s)|-1} \sum_{a_2 \in A(s),\, a_1 \neq a_2} \left(\hat{Q}(s, a_1) \cdot \hat{Q}(s, a_2)\right)$
- ( ) None of the above

Explanation

Pacman must select the best two actions (two maxes), and then flip a coin to determine which to perform (average their values).

Submit You have used 0 of 2 attempts

Generating Speech Output played within the problem

Part 5

0.0/3.0 points (graded)

Which of the following equations specify the relationship between V^* and $\hat{V}$ in general?

- $2V^*(s) = \hat{V}(s)$
- $V^*(s) = 2\hat{V}(s)$
- $(V^*(s))^2 = |\hat{V}(s)|$
- $|V^*(s)| = (\hat{V}(s))^2$
- $\frac{1}{|A(s)|}\sum_{a\in A(s)}\sum_{s'\in S} T(s,a,s')\hat{V}(s') = V^*(s)$
- $\frac{1}{|A(s)|}\sum_{a\in A(s)}\sum_{s'\in S} T(s,a,s')V^*(s') = \hat{V}(s)$
- $\frac{1}{|A(s)|}\sum_{a\in A(s)}\sum_{s'\in S} T(s,a,s')(R(s,a,s') + \gamma V^*(s')) = \hat{V}(s)$
- $\frac{1}{|A(s)|}\sum_{a\in A(s)}\sum_{s'\in S} T(s,a,s')(R(s,a,s') + \gamma \hat{V}(s')) = V^*(s)$
- None of the above. ✔

Explanation

None of the above are valid relationships between V^* and $\hat{V}$. It may be possible to construct MDPs where some of the above are satisfied, but they won't be satisfied for all MDPs.

Submit

You have used 0 of 2 attempts

Generating Speech Output

Answers are displayed within the problem

Generating Speech Output