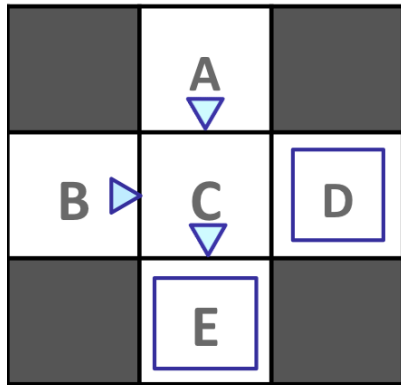


CS 4499/5599 HW 5

1 Model-Based RL: Grid

Input Policy π



Assume: $\gamma = 1$

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

What model would be learned from the above observed episodes?

$T(A, \text{south}, C) =$

$T(B, \text{east}, C) =$

$T(C, \text{south}, E) =$

$T(C, \text{south}, D) =$

2 Model-Based RL: Cycle

s	a	s'	r	s	a	s'	r	s	a	s'	r
A	Clockwise	B	0.0	B	Clockwise	A	-10.0	C	Clockwise	A	0.0
A	Clockwise	C	-1.0	B	Clockwise	C	0.0	C	Clockwise	A	0.0
A	Clockwise	C	-1.0	B	Clockwise	A	-10.0	C	Clockwise	A	0.0
A	Clockwise	B	0.0	B	Clockwise	C	0.0	C	Clockwise	A	0.0
A	Clockwise	B	0.0	B	Clockwise	C	0.0	C	Clockwise	A	0.0
A	Counterclockwise	C	0.0	B	Counterclockwise	C	0.0	C	Counterclockwise	B	0.0
A	Counterclockwise	C	0.0	B	Counterclockwise	C	0.0	C	Counterclockwise	B	0.0
A	Counterclockwise	C	0.0	B	Counterclockwise	A	-10.0	C	Counterclockwise	B	0.0
A	Counterclockwise	C	0.0	B	Counterclockwise	C	0.0	C	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0	B	Counterclockwise	A	-10.0	C	Counterclockwise	B	0.0

Part 1: We start by estimating the transition function, $T(s,a,s')$ and reward function $R(s,a,s')$ for this MDP. Fill in the missing values in the following table for $T(s,a,s')$ and $R(s,a,s')$. Discount Factor, $\gamma = 0.5$.

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
A	Clockwise	B	M	N
A	Clockwise	C	O	P
A	Counterclockwise	B	0.200	0.000
A	Counterclockwise	C	0.800	0.000
B	Clockwise	A	0.400	-10.000
B	Clockwise	C	0.600	0.000
B	Counterclockwise	A	0.400	-10.000
B	Counterclockwise	C	0.600	0.000
C	Clockwise	A	1.000	0.000
C	Counterclockwise	B	1.000	0.000

M=

N=

O=

P=

Part 2: Now we will run Q-iteration using the estimated T and R functions. The values of $Q_k(s,a)$, are given in the table below.

	A	B	C
Clockwise	-1.6	-4.0	0.0
Counterclockwise	-0.4	-4.0	-2.0

Fill in the values for $Q_{k+1}(s,a)$ for the state A.

	A
Clockwise	
Counterclockwise	

Part 3: Suppose Q-iteration converges to the following Q^* function $Q^*(s,a)$.

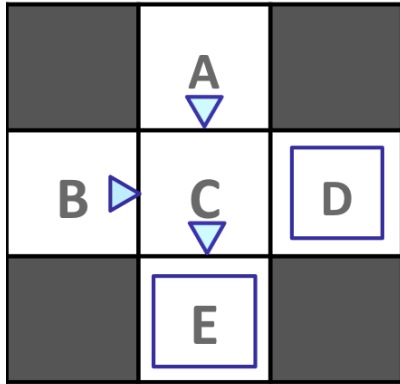
	A	B	C
Clockwise	-1.707	-4.183	-0.261
Counterclockwise	-0.523	-4.183	-2.092

What is the optimal action, either Clockwise or Counterclockwise, for each of the states?

A	B	C

3 Direct Evaluation

Input Policy π



Assume: $\gamma = 1$

Observed Episodes (Training)

Episode 1

A, south, C, -1
C, south, E, -1
E, exit, x, +10

Episode 2

B, east, C, -1
C, south, D, -1
D, exit, x, -10

Episode 3

B, east, C, -1
C, south, E, -1
E, exit, x, +10

Episode 4

A, south, C, -1
C, south, E, -1
E, exit, x, +10

What are the estimates for the following quantities as obtained by direct evaluation:

$$\hat{V}^{\pi}(A) =$$

$$\hat{V}^{\pi}(B) =$$

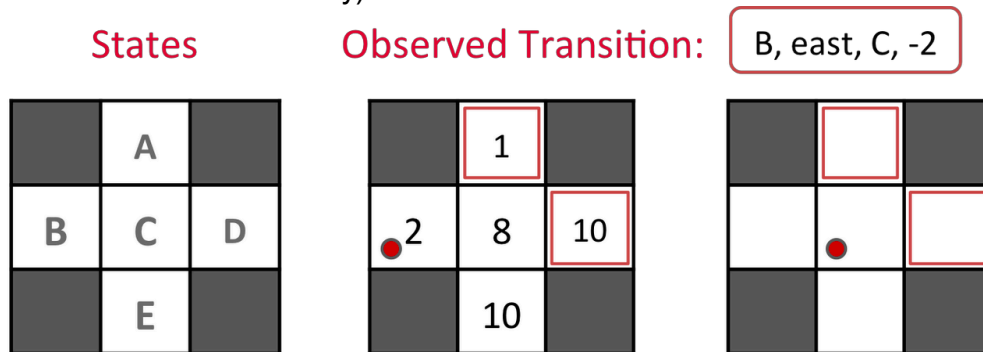
$$\hat{V}^{\pi}(C) =$$

$$\hat{V}^{\pi}(D) =$$

$$\hat{V}^{\pi}(E) =$$

4 Temporal Difference Learning

Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^π for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives a reward of -2. Assuming $\gamma=1$, $\alpha=1/2$, what are the value estimates after the TD learning update? (note: the value will change for one of the states only)



Assume: $\gamma = 1$, $\alpha = 1/2$

$$V^\pi(s) \leftarrow (1 - \alpha)V^\pi(s) + \alpha [R(s, \pi(s), s') + \gamma V^\pi(s')]$$

$$\hat{V}^\pi(A) =$$

$$\hat{V}^\pi(B) =$$

$$\hat{V}^\pi(C) =$$

$$\hat{V}^\pi(D) =$$

$$\hat{V}^\pi(E) =$$

5 Model-Free RL: Cycle

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning.

Assume, the discount factor, γ is 0.5 and the step size for Q-learning, α is 0.5.

Our current Q function, $Q(s,a)$, is as follows.

	A	B	C
Clockwise	0.172	0.75	-4.188
Counterclockwise	-1.0	-3.539	2.0

The agent encounters the following samples.

s	a	s'	r
A	Counterclockwise	C	-9.0
C	Counterclockwise	B	8.0

Process the samples given above. Below fill in the Q-values after both samples have been accounted for.

	A	B	C
Clockwise			
Counterclockwise			

6 Q-Learning Properties

Indicate whether the following assertions are **true** or **false** in general for Q-Learning to converge to the optimal Q-values.

- a) It is necessary that every state-action pair is visited infinitely often.
- b) It is necessary that the learning rate α (weight given to new samples) is decreased to over time.
- c) It is necessary that the discount γ is less than 0.5.
- d) It is necessary that actions get chosen according to $\operatorname{argmax}_a Q(s,a)$.

7 Exploration and Exploitation

For each of the following action-selection methods, indicate which option describes it best between **mostly exploration**, **mostly exploitation**, or **mix of both**.

- a) With probability p , select $\operatorname{argmax}_a Q(s,a)$. With probability $(1-p)$, select a random action. $p=0.99$.
- b) Select action a with the following probability, where τ is a temperature parameter that is decreased over time:

$$P(a|s) = \frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$$

- c) Always select a random action.
- d) Keep track of a count, $K_{s,a}$, for each state-action tuple, (s,a) , of the number of times that tuple has been seen and select $\operatorname{argmax}_a [Q(s,a) - K_{s,a}]$.

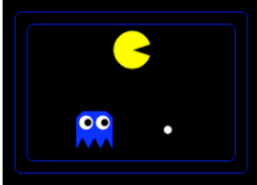
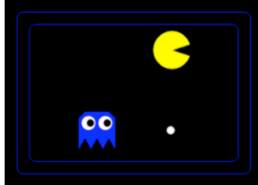
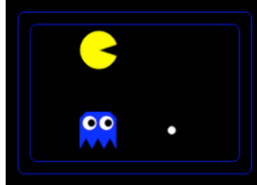
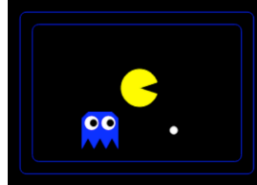
Which of the four method(s) would be advisable to use when doing Q-Learning?

8 Feature-Based Representation: Actions

Consider the Pacman board states presented below. The agent is considering possible actions to take; these are represented by the images. The agent is using feature-based representation to estimate the $Q(s,a)$ value of taking an action in a state, and the features the agent uses are:

- $f_1 = 1/(\text{Manhattan distance to closest food} + 1)$
- $f_2 = 1/(\text{Manhattan distance to closest ghost} + 1)$

For example, the feature representation $f(s = A, a = \text{STOP}) = [1/4, 1/4]$.

State	$a=\text{STOP}$	$a=\text{RIGHT}$	$a=\text{LEFT}$	$a=\text{DOWN}$
A				
$f(s, a)$	[0.25, 0.25]	[1/3, 0.2]	[0.2, 1/3]	[1/3, 1/3]

The agent picks the action according to

$$\operatorname{argmax}_a Q(s,a) = w^T f(s,a) = w_0 f_0(s,a) + w_1 f_1(s,a),$$

where the features $f_i(s,a)$ are as defined above, and w is a weight vector. Using the weight vector $w=[0.2,0.5]$, which action, of the ones shown above, would the agent take from state A?

Using the weight vector $w=[0.2,-1]$, which action, of the ones shown above, would the agent take from state A?

9 Feature-Based Representation: Update

Consider the following feature based representation of the Q-function:

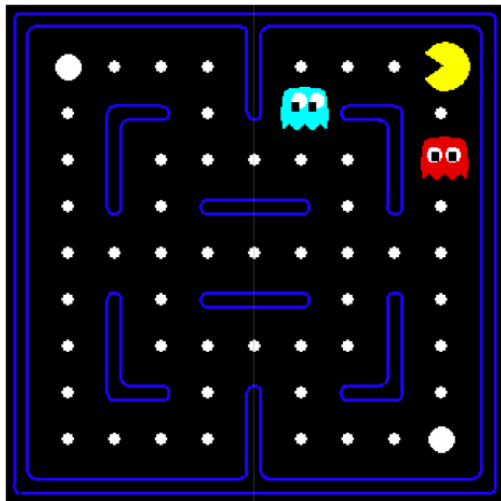
$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a)$$

with

$$f_1(s,a) = 1 / (\text{Manhattan distance to nearest dot after having executed action } a \text{ in state } s)$$

$$f_2(s,a) = (\text{Manhattan distance to nearest ghost after having executed action } a \text{ in state } s)$$

Part 1: Assume $w_1 = 1$, $w_2 = 10$. For the state s shown below, find the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot. Assume the ghosts do not move and that the distance between any two dots is 1.

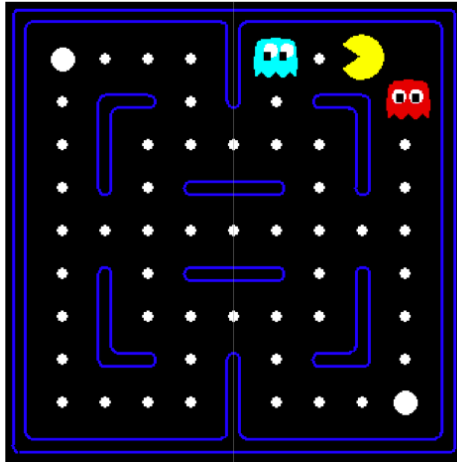


$$Q(s, \text{West}) =$$

$$Q(s, \text{South}) =$$

Based on this approximate Q-function, which action would be chosen?

Part 2: Assume Pac-Man moves West. This results in the state s' shown below.



The reward for this transition is $r = +10 - 1 = 9$ (+10: for food pellet eating, -1 for time passed). Fill in the following quantities. Assume that the red and blue ghosts are both sitting on top of a dot.

$$Q(s', \text{West}) =$$

$$Q(s', \text{East}) =$$

What is the sample value (assuming $\gamma = 1$)?

$$\text{sample} = [r + \gamma \max_{a'} Q(s', a')] =$$

Part 3: Now let's compute the update to the weights. Let $\alpha = 0.5$.

$$\text{difference} = [r + \gamma \max_{a'} Q(s', a')] - Q(s, a) =$$

$$w_1 \leftarrow w_1 + \alpha (\text{difference}) f_1(s, a) =$$

$$w_2 \leftarrow w_2 + \alpha (\text{difference}) f_2(s, a) =$$