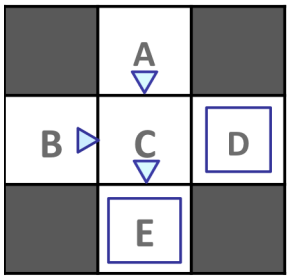
1. (5 points) Model-Based RL: Grid



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?

(a) T(A, south, C) = ***1***

(b) T(B, east, C) = ***1***

(c) T(C, south, E) = ***0.75***

(d) T(C, south, D) = ***0.25***

Explanation.

a) The action south is taken twice from state A, and both times results in state C. $\frac{2}{2} = 1$

b) The action east is taken twice from state B, and both times results in state C. $\frac{2}{2} = 1$

c) The action south is taken four times from state C, and results in state E three times. $\frac{3}{4} = 0.75$

d) The action south is taken four times from state C, and results in state D one time. $\frac{1}{4} = 0.25$

2. (22 points) Model-Based 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 samples of what an agent 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, we will first estimate the model (the transition function and the reward function), and then use the estimated model to find the optimal actions.

To find the optimal actions, model-based RL proceeds by computing the optimal V or Q value function with respect to the estimated T and R. This could be done with value iteration, or policy iteration, or Q-value iteration. In this problem we will use Q-value iteration.

Consider the following samples that the agent encountered.

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	0.400	0.000
A	Counterclockwise	C	0.600	-8.000
B	Clockwise	A	0.800	-3.000
B	Clockwise	C	0.200	0.000
B	Counterclockwise	A	0.800	-10.000
B	Counterclockwise	C	0.200	0.000
C	Clockwise	A	0.600	0.000
C	Clockwise	B	0.400	6.000
C	Counterclockwise	A	0.200	0.000
C	Counterclockwise	B	0.800	-8.000

Figure 1: Table for T(s, a, s') and R(s, a, s').

s	a	s'	r
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	B	0.0
A	Clockwise	C	-10.0
A	Clockwise	C	-10.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	C	-8.0
A	Counterclockwise	B	0.0
A	Counterclockwise	B	0.0
A	Counterclockwise	C	-8.0

s	a	s'	r
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	A	-3.0
B	Clockwise	C	0.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	A	-10.0
B	Counterclockwise	C	0.0

s	a	s'	r
C	Clockwise	A	0.0
C	Clockwise	B	6.0
C	Clockwise	B	6.0
C	Clockwise	A	0.0
C	Clockwise	A	0.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	B	-8.0
C	Counterclockwise	A	0.0
C	Counterclockwise	B	-8.0

(a) 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 table for T(s, a, s') and R(s, a, s') shown in Figure 1.

M = ***0.6*** N = ***0*** O = ***0.4*** P = ***-10***

Explanation.

M. Clockwise action taken 5 times from A, and went to B three times, $3/5 = 0.6$.

N. transition (A,clockwise,B) occurred 3 times with reward 0 each time, $(0 + 0 + 0)/3 = 0$.

O. Clockwise action taken 5 times from A, and went to C two times, $2/5 = 0.4$.

P. Both occurrences of (A,clockwise,C) had reward -10.

(b) Run Q-iteration using the estimated T and R functions above. The values of $Q_k(s, a)$, are given in the table below.

	A	B	C
Clockwise	-4.24	-3.76	0.72
Counterclockwise	-4.56	-9.36	-7.76

Compute $Q_{k+1}(s, a)$ in each of the following cases. Write your answer on the line provided.

i. Q_{k+1}(A, clockwise)
Explanation. Let cw = clockwise and ccw = counterclockwise.
$V_k(B) = \max(Q_k(B,cw), Q_k(B,ccw)) = -3.76$
$V_k(C) = \max(Q_k(C,cw), Q_k(C,ccw)) = 0.72$
$Q_{k+1}(A, cw) = T(A,cw,B) \cdot (R(A,cw,B) + \gamma\ V_k(B))$
$+ T(A,cw,C) \cdot (R(A,cw,C) + \gamma\ V_k(C))$
$= 0.6 \cdot (0 + .5 \cdot -3.76) + 0.4 \cdot (-10 + 0.5 \cdot 0.72) = -4.984$

i. ***-4.984***

ii. Q_{k+1}(A, counterclockwise)
Explanation. Let ccw = counterclockwise.
$Q_{k+1}(A, ccw) = T(A,ccw,B) \cdot (R(A,ccw,B) + \gamma\ V_k(B))$
$+ T(A,ccw,C) \cdot (R(A,ccw,C) + \gamma\ V_k(C))$
$= 0.4 \cdot (0 + 0.5 \cdot -3.76) + .6 \cdot (-8 + 0.5 \cdot 0.72) = -5.336$

ii. ***-5.336***

iii. Q_{k+1}(B, clockwise)
Explanation. Let cw = clockwise; ccw = counterclockwise.
$V_k(A) = \max(Q_k(A,cw),Q_k(A,ccw)) = -4.24$
$Q_{k+1}(B, cw) = T(B,cw,A) \cdot (R(B,cw,A) + \gamma\ V_k(A))$
$+ T(B,cw,C) \cdot (R(B,cw,C) + \gamma\ V_k(C))$
$= 0.8 \cdot (-3 + 0.5 \cdot -4.24) + 0.2 \cdot (0 + .5 \cdot 0.72) = -4.024$

iii. ***-4.024***

iv. Q_{k+1}(B, counterclockwise)
Explanation. Let ccw = counterclockwise.
$Q_{k+1}(B, ccw) = T(B,ccw,A) \cdot (R(B,ccw,A) + \gamma\ V_k(A))$
$+ T(B,ccw,C) \cdot (R(B,ccw,C) + \gamma\ V_k(C))$
$= 0.8 \cdot (-10 + 0.5 \cdot -4.24) + 0.2 \cdot (0 + 0.5 \cdot .72) = -9.624$

iv. ***-9.624***

v. Q_{k+1}(C, clockwise)
Explanation. Let cw = clockwise; ccw = counterclockwise.
$Q_{k+1}(C, cw) = T(C,cw,A) \cdot (R(C,cw,A) + \gamma\ V_k(A))$
$+ T(C, cw, B) \cdot (R(C, cw, B) + \gamma\ V_k(B))$
$= 0.6 \cdot (0 + 0.5 \cdot -4.24) + 0.4 \cdot (6 + 0.5 \cdot -3.76) = 0.376$

v. ***0.376***

vi. Q_{k+1}(C, counterclockwise)
Explanation. Let ccw = counterclockwise.
$Q_{k+1}(C, ccw) = T(C, ccw, A) \cdot (R(C, ccw, A) + \gamma\ V_k(A))$
$+ T(C, ccw, B) \cdot (R(C, ccw, B) + \gamma\ V_k(B))$
$= 0.2 \cdot (0 + 0.5 \cdot -4.24) + 0.8 \cdot (-8 + 0.5 \cdot -3.76) = -8.328$

vi. ***-8.328***

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

	A	B	C
Clockwise	-5.399	-4.573	-0.134
Counterclockwise	-5.755	-10.173	-8.769

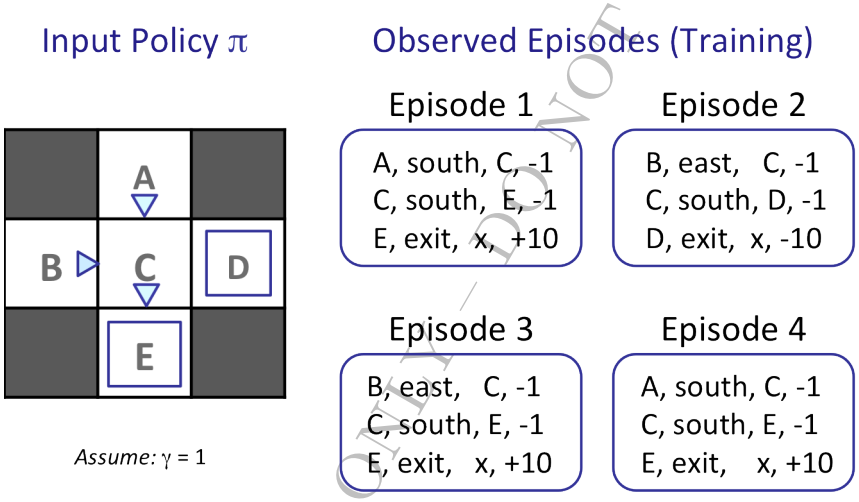
What is the optimal action, either clockwise or counterclockwise, for each of the states?

A: √ ***clockwise*** □ counterclockwise
Explanation. We select the action with the highest Q value: $-5.399 > -5.755$.

B: √ ***clockwise*** □ counterclockwise Explanation. $-4.573 > -10.173$

C: √ ***clockwise*** □ counterclockwise Explanation. $-0.134 > -8.769$

3. (10 points) Direct Evaluation



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

$\hat{V}^\pi(\text{A}) =$ ***8*** $\hat{V}^\pi(\text{B}) =$ ***-2*** $\hat{V}^\pi(\text{C}) =$ ***4***

$\hat{V}^\pi(\text{D}) =$ ***-10*** $\hat{V}^\pi(\text{E}) =$ ***10***

Explanation.

The estimated value of $\hat{V}^\pi(s)$ is equal to the average value achieved starting from that state.

$\hat{V}^\pi(\text{A})$: Episodes 1 and 4 start from state A; both result in a utility of 8, yielding $(8+8)/2 = 8$

$\hat{V}^\pi(\text{B})$: Episodes 2 and 3 start from B; they result in -12 and 8, respectively, yielding $(8\text{-}12)/2 = \text{-}2$.

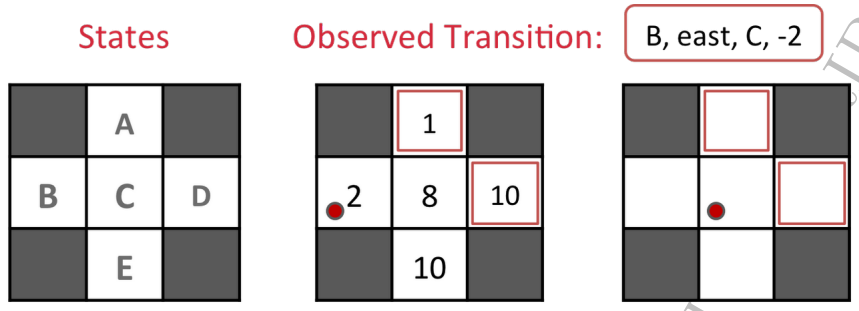
$\hat{V}^\pi(\text{C})$: State C is visited in every episode. The remaining rewards from C in episodes 1, 3, and 4 total 9, while those in episode 2 total -11, yielding $(9+9+9\text{-}11)/4 = 4$.

$\hat{V}^\pi(\text{D})$: State D is only visited in episode 2 and has remaining utility -10.

$\hat{V}^\pi(\text{E})$: State E is visited in episodes 1, 3, and 4 and has remaining utility 10 in each state, yielding $(10 + 10 + 10)/3 = 10$.

4. (10 points) 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: γ = 1, α = 1/2

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

$\hat{V}^\pi$(A) = ***1*** $\hat{V}^\pi$(B) = ***4*** $\hat{V}^\pi$(C) = ***8***

$\hat{V}^\pi$(D) = ***10*** $\hat{V}^\pi$(E) = ***10***

Explanation. The only value that gets updated is $\hat{V}^\pi(B)$, because the only transition observed starts in state B, and $\hat{V}^\pi(\text{B}) = 0.5 \cdot 2 + 0.5 \cdot (-2 + 8) = 4$.

5. (12 points) 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 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	1.501	-0.451	2.73
Counterclockwise	3.153	-6.055	2.133

The agent encounters the following samples.

s	a	s'	r
A	Counterclockwise	C	8.0
C	Counterclockwise	A	0.0

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

Q(A, clockwise) = ***1.501***

Q(A, counterclockwise) = ***6.259***

Q(B, clockwise) = ***-0.451***

Q(B, counterclockwise) = ***-6.055***

Q(C, clockwise) = ***2.73***

Q(C, counterclockwise) = ***2.631***

Explanation. For each s, a, s', r transition sample, we update the Q value function as follows:

$Q(s, a) = (1\text{-}\alpha)\ Q(s, a) + \alpha\ (R(s, a, s') + \gamma \max \{ Q(s', a') : a' \in \text{Actions}\}$.

First we update

$Q(\text{A, counterclockwise}) = 0.5 \cdot 3.153 + 0.5 \cdot (8 + 0.5 \cdot 2.73) = 6.259$.

Then we update

$Q(\text{C, counterclockwise}) = 0.5 \cdot 2.133 + 0.5 \cdot (0 + 0.5 \cdot 6.259) \approx 2.631$

using the updated value of Q(A, counterclockwise), 6.259, in this calculation (instead of the maximum of the original Q values for state C).

Finally, note that the other four values stay the same since there are only two samples.

6. (5 points) Q-Learning Properties

In general, for Q-Learning to converge to the optimal Q-values...

- √ ***It is necessary that every state-action pair is visited infinitely often.***
- √ ***It is necessary that the learning rate α (weight given to new samples) is decreased to 0 over time.***
- □ It is necessary that the discount γ is less than 1/2.
- □ It is necessary that actions get chosen according to $\text{argmax}_a\ Q(s, a)$.

Explanation.

a) In order to ensure convergence in general for Q learning, this has to be true. In practice, we generally care about the policy, which converges well before the values do, so it is not necessary to run it infinitely often.

b) In order to ensure convergence in general for Q learning, this has to be true.

c) The discount factor must be greater than 0 and less than 1, not 0.5.

d) This would actually do rather poorly, because it is purely exploiting based on the Q-values learned thus far, and not exploring other states to try and find a better policy.

7. (12 points) Exploration and Exploitation

(a) For each of the following action-selection methods, indicate which option describes it best.

i. With probability p, select argmax$_a$ Q(s, a). With probability 1-p, select a random action. p = 0.99.

□ Mostly exploration √ ***Mostly exploitation*** □ Mix of both

Explanation. 99% of the time, it will choose the greedy action with respect to the current Q values, which is exploitation, so this is mostly exploitation.

ii. Select action a with probability

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

where τ is a temperature parameter that is decreased over time.

□ Mostly exploration □ Mostly exploitation √ ***Mix of both***

Explanation. When τ is high this method results in mostly exploration, and as τ is decreased, it results in more exploitation. Consider when τ is ∞, in this case all actions are selected uniformly at random. Then, as τ decreases, actions with higher Q values start to get selected with higher probability resulting in more exploitation.

iii. Always select a random action.

√ ***Mostly exploration*** □ Mostly exploitation □ Mix of both

Explanation. By not considering the policy at all, it is doing no exploitation, and is thus mostly exploration.

iv. Keep track of a count, K$_{sa}$, for each state-action tuple, (s, a), of the number of times that tuple has been seen and select argmax$_a$ (Q(s, a) - K$_{sa}$).

□ Mostly exploration □ Mostly exploitation √ ***Mix of both***

Explanation. This method initially does mostly exploitation, but as the K$_{sa}$ counts increase for commonly seen tuples, it begins trying new tuples with lower Q values, which is more exploration.

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

□ i. √ ***ii.*** □ iii. √ ***iv.***

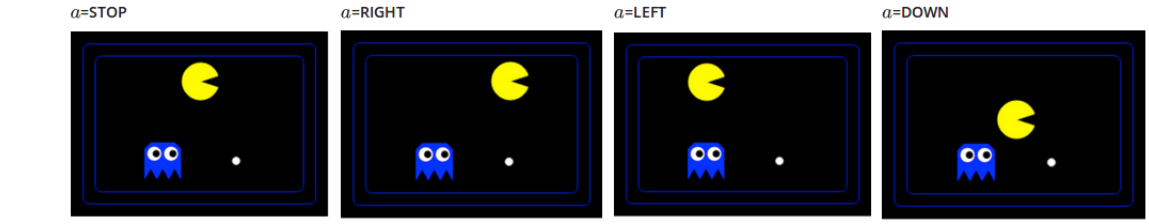
Explanation. In general, it is best to use methods that mix exploration and exploitation when doing Q-learning.

8. (6 points) Feature-Based Representation: Actions

A Pacman agent is using a 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_0 = 1/(Manhattan distance to closest food + 1)
f_1 = 1/(Manhattan distance to closest ghost + 1)

The images below show the result of taking actions STOP, RIGHT, LEFT, and DOWN from a state A. The feature vectors for each action are shown below each image. For example, the feature representation f(s=A, a=STOP) = [1/4, 1/4].



$f(s,a)$	[0.25, 0.25]	[1/3, 0.2]	[0.2, 1/3]	[1/3, 1/3]

The agent picks the action according to:

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.

(a) Using the weight vector w = [0.2, 0.5], which action, of the ones shown above, would the agent take from state A?

☐ STOP ☐ RIGHT ☐ LEFT √ ***DOWN***

Explanation. STOP: 0.2 · 0.25 + 0.5 · 0.25 = 0.175; RIGHT: 0.2 · 0.33 + 0.5 · 0.2 = 0.166; LEFT: 0.2 · 0.2 + 0.5 · 0.33 = 0.205; DOWN: 0.2 · 0.33 + 0.5 · 0.33 = 0.231. Since 0.231 is the highest value, the agent would take the DOWN action.

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

☐ STOP √ ***RIGHT*** ☐ LEFT ☐ DOWN

Explanation. STOP: 0.2 · 0.25-0.25 = -0.2; RIGHT: 0.2 · 0.33-0.2 = -0.134; LEFT: 0.2 · 0.2-0.33 = -0.29; DOWN: 0.2 · 0.33-0.33 = -0.264. Since -0.134 is the highest value, the agent would take the RIGHT action.

9. (18 points) 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

$f_1(s, a) = 1$ / (Manhattan distance to nearest dot after executing action a in state s)

$f_2(s, a) =$ (Manhattan distance to nearest ghost after executing action a in state s)

(a) 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.

$Q(s, \text{West}) =$ ***31*** $Q(s, \text{South}) =$ ***11***

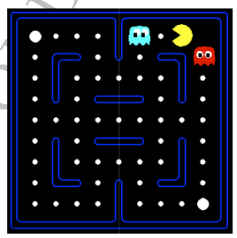
Based on this approximate Q-function, which action would be chosen:

√ ***West*** □ South

Explanation.

$Q(s, \text{West}) = 1 \cdot 1 + 10 \cdot 3 = 31$, and $Q(s, \text{South}) = 1 \cdot 1 + 10 \cdot 1 = 11$, so West is best.

(b) Assume Pac-Man moves West. This results in the state s' shown below. Pac-Man receives reward 9 (10 for eating a dot and -1 living penalty).



$Q(s', \text{West}) =$ ***11*** $Q(s', \text{East}) =$ ***11***

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

$\text{sample} = r + \gamma \max\{Q(s', a') : a' \in \text{Actions}\} =$ ***20***

Explanation.

$Q(s', \text{West}) = 1 \cdot 1 + 10 \cdot 1 = 11$; $Q(s', \text{East}) = 1 \cdot 1 + 10 \cdot 1 = 11$; sample $= 9 + 1 \cdot 11 = 20$

(c) Now let's compute the update to the weights. Let $\alpha = 0.5$.

$\text{difference} = (r + \gamma \max\{Q(s', a') : a' \in \text{Actions}\}) - Q(s, a) =$ ***-11***

$w_1 \leftarrow w_1 + \alpha\ (\text{difference})\ f_1(s, a) =$ ***-4.5***

$w_2 \leftarrow w_2 + \alpha\ (\text{difference})\ f_2(s, a) =$ ***-6.5***

Explanation.

difference = 20 - 31 = -11; $w_1 = 1 + 0.5 \cdot -11 \cdot 1 = -4.5$; $w_2 = 10 + 0.5 \cdot -11 \cdot 3 = -6.5$