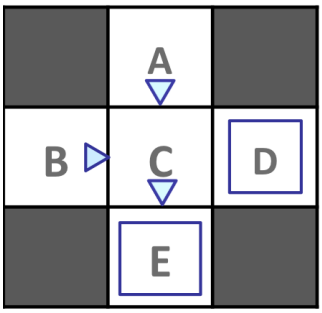
Q1 Model-Based RL: Grid

5 Points

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, south, C) =

1

T(B, east, C) =

1

T(C, south, E) =

.75

T(C, south, D) =

.25

Q2 Model-Based RL: Cycle

22 Points

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

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 any of value iteration, policy iteration, or Q-value iteration. Last week you already solved some exercises that involved value iteration and policy iteration, so we will go with Q value iteration in this exercise.

Consider the following samples that the agent encountered.

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

Q2.1
8 Points

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	0.400	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

M

.6

N

0

O

.4

P

-10

Q2.2
8 Points

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	-4.24	-3.76	0.72
Counterclockwise	-4.56	-9.36	-7.76

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

Q(A, clockwise)

-4.98

Q(A, counterclockwise)

-5.336

Q(B, clockwise)

-4.024

Q(B, counterclockwise)

-9.624

Q(C, clockwise)

.376

Q(C, counterclockwise)

-8.328

Q2.3
6 Points

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

B

Clockwise

Counterclockwise

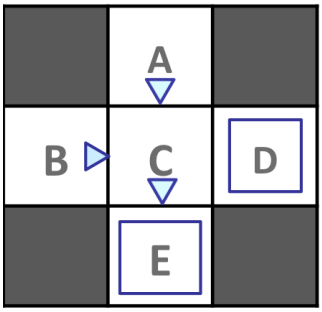
C

Clockwise

Counterclockwise

Q3 Direct Evaluation
10 Points

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:

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

8

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

-2

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

4

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

-10

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

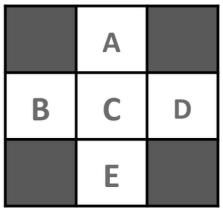
10

Q4 Temporal Difference Learning

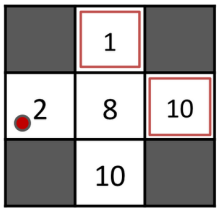
10 Points

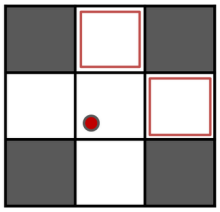
Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function V^π for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives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

States



Observed Transition: B, east, C, -2





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

Q5 Model-Free RL: Cycle

12 Points

We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.

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	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)

-.451

Q(B, counterclockwise)

-6.055

Q(C, clockwise)

2.73

Q(C, counterclockwise)

2.63125

Q6 Q-Learning Properties

5 Points

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

- [x] It is necessary that every state-action pair is visited infinitely often.
- [x] 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 0.5.
- [ ] It is necessary that actions get chosen according to $\arg\max_a Q(s, a)$.

Q7 Exploration and Exploitation
12 Points

Q7.1
8 Points

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

A: 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

B: 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

C: Always select a random action.

Mostly exploration

Mostly exploitation

Mix of both

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 $argmax_a[Q(s,a) - K_{s,a}]$.

Mostly exploration

Mostly exploitation

Mix of both

Q7.2
4 Points

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

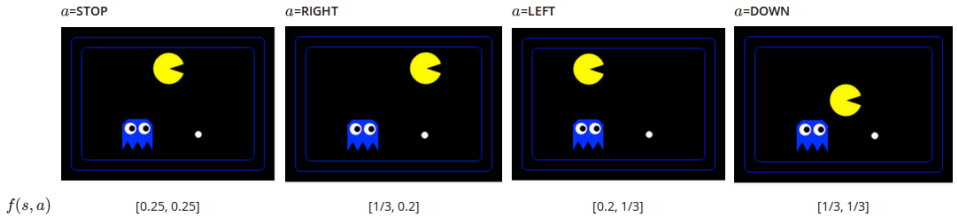
- [ ] A
- [x] B
- [ ] C
- [x] D

Q8 Feature-Based Representation: Actions
6 Points

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 $f(s = A, a = \textbf{STOP}) = [1/4, 1/4]$.



The agent picks the action according to $\arg\max_a Q(s, a) = w^T f(s, a) = w_0 f_0(s, a) + w_1 f_1(s, a)$, 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?

STOP

RIGHT

LEFT

DOWN

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

Q9 Feature-Based Representation: Update

18 Points

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

$f_1(s,a) =$

$1/(\text{Manhattan distance to nearest dot after having executed action } a \text{ in}$

$f_2(s,a) =$

$(\text{Manhattan distance to nearest ghost after having executed action } a \text{ in}$

Q9.1
6 Points

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.

$Q(s, West) =$

31

$Q(s, South) =$

11

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

West

South

Q9.2
6 Points

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).

$Q(s', West) =$

11

$Q(s', East) =$

11

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

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

20

Q9.3
6 Points

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

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

-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

HW 5 (Electronic Component)

● Graded

Student
ریحانه شاهرخیان

Total Points
100 / 100 pts

Question 1
Model-Based RL: Grid — **5** / 5 pts

Question 2
Model-Based RL: Cycle — **22** / 22 pts

2.1 (no title) — **8** / 8 pts

2.2 (no title) — **8** / 8 pts

2.3 (no title) — **6** / 6 pts

Question 3

Direct Evaluation **10** / 10 pts

Question 4

Temporal Difference Learning **10** / 10 pts

Question 5

Model-Free RL: Cycle **12** / 12 pts

Question 6

Q-Learning Properties **5** / 5 pts

Question 7

Exploration and Exploitation **12** / 12 pts

7.1 (no title) **8** / 8 pts

7.2 (no title) **4** / 4 pts

Question 8

Feature-Based Representation: Actions **6** / 6 pts

Question 9

Feature-Based Representation: Update **18** / 18 pts

9.1 (no title) **6** / 6 pts

9.2 (no title) **6** / 6 pts

9.3 (no title) **6** / 6 pts