**0/40** Questions Answered

# HW 10 (Electronic Component)

**STUDENT NAME**

Search students by name or email...

## **Q1** Local Optima and Gradient Descent

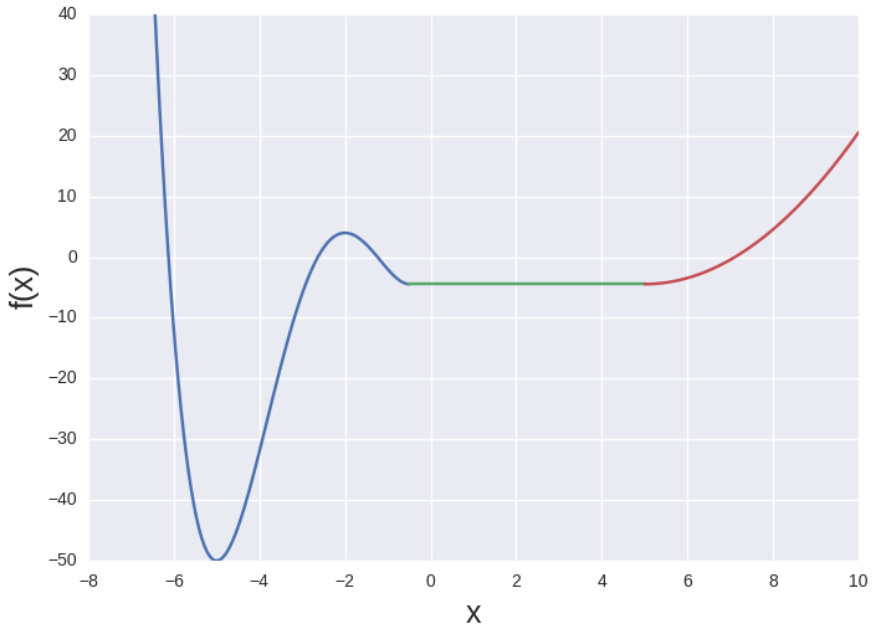
16 Points

After a busy year of chasing ghosts, Pacman and Paclady are planning to visit the Kakslauttanen Arctic Resort for their winter vacation. Paclady who is particularly fond of skiing, excitedly begins planning ahead. Pacman, who is apprehensive of skiing (when asked why, he rambles on about the Aspen Red Ghost Chase of 2012, but we won't get into that), reluctantly agreed to go skiing, but under one condition: Paclady must tell Pacman how steep the slopes are at several points of interest.

Paclady asks the resort for terrain details, and receives the following graph. The resort says at any given location $x$, $f(x)$ models the terrain height. Specifically:

- when $x \leq -\frac{1}{2}$, $f(x) = \frac{1}{2}x^4 + 5x^3 + \frac{27}{2}x^2 + 10x$,
- when $-\frac{1}{2} \leq x \leq 5$, $f(x) = -\frac{71}{16}$,
- and when $x \geq 5$, $f(x) = x^2 - 10x + \frac{329}{16}$.

See below for a plot:




The local optima for $f$ lie at $x = -5$ and $x = -2$, with a plateau in the region $-1/2 \leq x \leq 5$.

## Q1.1

2 Points

Paclady decides to compute derivatives to measure how steep slopes are.

Evaluate $f'(-6)$.

Enter your answer here

Save Answer

## Q1.2

2 Points

Evaluate $f'(0)$.

Enter your answer here

Save Answer

## Q1.3
2 Points

Evaluate $f'(8)$.

Enter your answer here

Save Answer

## Q1.4
2 Points

Pacman and Paclady get to the resort, and have a fantastic time skiing, but get lost. Unfortunately, a blizzard kicks in right then, reducing visibility. As Pacman panics and brings up the Aspen Red Ghost Chase of 2012, Paclady remembers that their glass igloo cabin is located at the global minimum elevation point of the resort ($x = -5$). The blizzard complicates things, since they can't ski due to the reduced visibility for safety.

After thinking for a minute, Pacman says, "Aha! We can get home in that case by following gradient descent, as long as we employ a small step size -- once we hit a gradient of $0$, we know we're home!" Paclady pauses and says, "Your algorithm almost works, but it depends on where in the resort we currently are."

Check all regions where Pacman and Paclady can be, and still find their igloo, assuming that they employ gradient descent with a small step size and stop walking when they encounter a gradient of $0$.

- [ ] $x < -5$
- [ ] $-5 < x < -2$
- [ ] $-2 < x < -1/2$
- [ ] $-1/2 < x < 3$
- [ ] $3 < x < 5$
- [ ] $x > 5$

Note: Make sure you select all of the correct options--there may be more than one!

Save Answer

## Q1.5

2 Points

While slowly trudging to their igloo via gradient descent, Pacman and Paclady get into an argument. Pacman complains that trudging down a hill is tiresome, and that they instead should have gotten an igloo closer to $x = 3$. Paclady says that Pacman's previous gradient descent algorithm wouldn't lead them to the igloo in this case, unless they were already at the igloo. Why is this the case?

- Gradient descent would cause Pacman and Paclady to reach $x = -2$ rather than $x = 3$, since it is at a local maximum.
- Gradient descent terminates when it reaches a gradient of $0$, and neighboring regions around $x = 3$ all have a gradient of $0$, so Pacman and Paclady would stop searching outside of $x = 3$, within the plateau.
- When gradient descent is stuck in a plateau, it searches for regions with negative rather than zero gradient.
- When gradient descent is stuck in a plateau, it searches for regions with positive rather than zero gradient.
- Gradient descent seeks to maximize a function, which would lead Pacman and Paclady either to $-\infty$ or to $\infty$.

Save Answer

## Q1.6
2 Points

In the following, we introduce Newton's method for optimization. The update rule is as follows:

$$x^{(t+1)} = x^{(t)} - \frac{f'(x^{(t)})}{f''(x^{(t)})}$$

With this update rule, when $x^{(t)} = -6$, what is $x^{(t+1)}$?

Enter your answer here

Save Answer

## Q1.7
2 Points

Following part 6, when $x^{(t)} = 8$, what is $x^{(t+1)}$?

Enter your answer here

Save Answer

## Q1.8
2 Points

For which of the following functions, Newton's method converges in one step?

- [ ] $f(x) = \frac{1}{2}x^4 + 5x^3 + \frac{27}{2}x^2 + 10x$
- [ ] $f(x) = x^2 - 10x + \frac{329}{16}$
- [ ] $f(x) = 5x^3 + \frac{27}{2}x^2 + 10x$
- [ ] $f(x) = \frac{27}{2}x^2 + 10x$
- [ ] $f(x) = \frac{1}{2}x^4 + 5x^3 + \frac{27}{2}x^2$
- [ ] $f(x) = \frac{1}{2}x^4 + 5x^3$

Save Answer

# Q2 Learning Rates
6 Points

Video Link
http://www.youtube.com/watch?v=FaDgovU4_0o

Watch the above YouTube video. There are three sub-panels. We will refer to the left one by (A), the middle one by (B), and the right one by (C). The same objective is being optimized by gradient descent, but has different learning rates in each subpanel.

## Q2.1
2 Points

Which animation corresponds to the lowest learning rate?

- Animation (A)
- Animation (B)
- Animation (C)

Save Answer

## Q2.2
2 Points

Which animation corresponds to the medium learning rate?

- Animation (A)
- Animation (B)
- Animation (C)

Save Answer

## Q2.3
2 Points

Which animation corresponds to the largest learning rate?

- Animation (A)
- Animation (B)
- Animation (C)

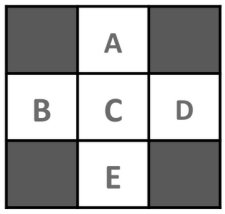
Save Answer

## **Q3** Temporal Difference Learning

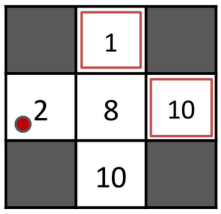
10 Points

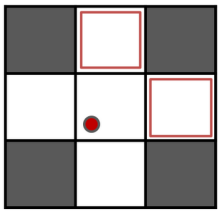
Consider the gridworld shown below. The left panel shows the name of each state A through E. The middle panel shows the current estimate of the value function $V^\pi$ for each state. A transition is observed, that takes the agent from state B through taking action east into state C, and the agent receives 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

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

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

Enter your answer here

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

Enter your answer here

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

Enter your answer here

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

Enter your answer here

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

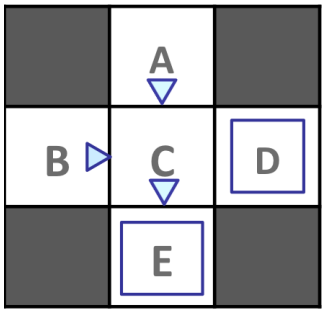
Enter your answer here

Save Answer

## **Q4** Model-Based RL: Grid

8 Points

Input Policy $\pi$




*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?

### **Q4.1**

2 Points

T(A, south, C) =

Enter your answer here

Save Answer

## Q4.2
2 Points

T(B, east, C) =

Enter your answer here

Save Answer

## Q4.3
2 Points

T(C, south, E) =

Enter your answer here

Save Answer

## Q4.4
2 Points

T(C, south, D) =

Enter your answer here

Save Answer

# Q5 Model-Based RL: Cycle
26 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 |

## Q5.1

2 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

Enter your answer here

Save Answer

## Q5.2

2 Points

N (following part 1)

Enter your answer here

Save Answer

## Q5.3
2 Points

O (following part 1)

Enter your answer here

Save Answer

## Q5.4
2 Points

P (following part 1)

Enter your answer here

Save Answer

## Q5.5
2 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)

Enter your answer here

Save Answer

Save Answer

## Q5.6
2 Points

Q(A, counterclockwise) (following part 5)

Enter your answer here

Save Answer

## Q5.7
2 Points

Q(B, clockwise) (following part 5)

Enter your answer here

Save Answer

## Q5.8
2 Points

Q(B, counterclockwise) (following part 5)

Enter your answer here

Save Answer

## Q5.9
2 Points

Q(C, clockwise) (following part 5)

Enter your answer here

Save Answer

## Q5.10
2 Points

Q(C, counterclockwise) (following part 5)

Enter your answer here

Save Answer

## Q5.11
2 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

Save Answer

## Q5.12
2 Points

B (following part 11)

- ○ Clockwise
- ○ Counterclockwise

Save Answer

## Q5.13
2 Points

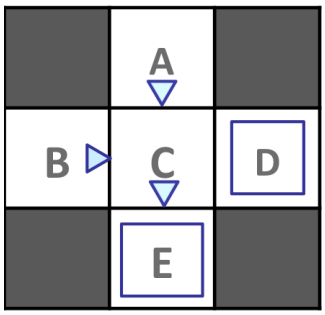
C (following part 11)

- ○ Clockwise
- ○ Counterclockwise

Save Answer

## Q6 Direct Evaluation
10 Points

**Input Policy π**




*Assume:* γ = 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:

## Q6.1
2 Points

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

Enter your answer here

Save Answer

## Q6.2
2 Points

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

Enter your answer here

Save Answer

## Q6.3
2 Points

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

Enter your answer here

Save Answer

## Q6.4
2 Points

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

Enter your answer here

Save Answer

## Q6.5
2 Points

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

Enter your answer here

Save Answer

# Q7 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, $\gamma$ is 0.5 and the step size for Q-learning, $\alpha$ 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.

## Q7.1
2 Points

Q(A, clockwise)

Enter your answer here

Save Answer

## Q7.2
2 Points

Q(A, counterclockwise)

Enter your answer here

Save Answer

## Q7.3
2 Points

Q(B, clockwise)

Enter your answer here

Save Answer

## Q7.4
2 Points

Q(B, counterclockwise)

Enter your answer here

Save Answer

## Q7.5
2 Points

Q(C, clockwise)

Enter your answer here

Save Answer

## Q7.6
2 Points

Q(C, counterclockwise)

Enter your answer here

Save Answer

Save All Answers

Submit & View Submission