# CS 486/686 Assignment 4 (84 marks in total)

Instructor: Alice Gao

Due Date: 6:00 pm on Tuesday, July 30, 2019

## Instructions

- Submit the assignment in the Dropbox labeled Assignment 4 Submissions in the Assignment 4 folder on LEARN. No late assignment will be accepted. This assignment is to be done individually.
- For any programming question, you may use the language of your choice. We highly recommend using Python. If you don't know Python, it may be worthwhile to learn it for this course.
- Lead TAs:
  - Atrisha Sarkar (a9sarkar@uwaterloo.ca)
  - Alexandre Parmentier (aparment@uwaterloo.ca)

  The TAs' office hours will be posted on the course website.
- Submit two files with the following naming conventions. **We will deduct 10 marks if you do not follow these conventions.**
  - **writeup.pdf**
    - Include your name, email address and student ID in the writeup file.
    - If you hand-write your solutions, make sure your handwriting is legible and take good quality pictures. You may get a mark of 0 if we cannot read your handwriting.
  - **code.zip**
    - Include your program, a script to run your program, and a README.txt file with instructions to run the script. You may get a mark of 0 if we cannot run your program.

# Learning goals

**Markov Decision Process**

- Trace the execution of and implement the value iteration algorithm to solve a Markov decision process.

# 1 Markov Decision Processes (84 marks)

In this question, we will continue to study the grid world discussed in class. We will vary the reward function $R(s)$ for any state $s \neq s_{24}$ and $s \neq s_{34}$ and investigate how the optimal policy changes as $R(s)$ changes.

Assume that $R(s_{24}) = -1$, $R(s_{34}) = +1$, and the discount factor $\gamma = 1$. The transition model stays the same: the agent moves in the intended direction with probability 0.8, moves to the left of the intended direction with probability 0.1, and moves to the right of the intended direction with probability 0.1.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | X | | -1 |
| 3 | | | | +1 |

1. Consider $R(s) = -1.0, \forall s \neq s_{24} \wedge s \neq s_{34}$. In iteration 0, let the true utility of every state to be 0. Implement the value iteration algorithm. Execute the value iteration algorithm until the true utilities of all the states have converged. Each true utility has converged when it stops changing up to three decimal places.

   Please show the optimal policy in a table format. See the table below for an example. $s_{23}$ shows a case when there is a tie between multiple actions for the optimal policy.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | $\uparrow$ | $\downarrow$ | $\leftarrow$ | $\rightarrow$ |
| 2 | $\uparrow$ | X | $\downarrow\uparrow$ | -1 |
| 3 | $\leftarrow$ | $\rightarrow$ | $\uparrow$ | +1 |

   **What to submit:**

- Show the optimal policy of the MDP. If there is a tie between multiple actions, show all of them as the optimal policy for a state.
- Show the true utilities of all the states until the iteration when all the true utilities have stopped changing up to three decimal places. That is, execute the algorithm until the earlier iteration $K$ such that the true utilities for all the states are the same in iteration $K-1$ and iteration $K$.
- A program you used to execute the value iteration algorithm and produce the output.

**Marking Scheme:** (36 marks)

- (6 marks) The TA can run your program to produce the specified output.
- (24 marks) The output shows the correct true utilities of all the states.
- (6 marks) The optimal policy is correct.

2. Determine how the optimal policy changes when $R(s)$ changes from $-1.6$ to $-0.5$.

   You only need to consider the values of $R(s)$ up to one decimal place. That is, you only need to consider $R(s) = -1.6, -1.5, -1.4, -1.3, -1.2, -1.1, -1.0, -0.9, -0.8, -0.7, -0.6, -0.5$.

   **What to submit:**

   - Divide up the values of $R(s)$ into ranges such that the optimal policy for each range of values remains the same.
   - State each range of values and state the optimal policy. If there is a tie between multiple actions, show all of them as the optimal policy for a state.
   - Indicate how the optimal policy changes from the last value in one range to the first value in the next range.

   **Marking Scheme:** (24 marks)

   - (6 marks) Correct ranges of values
   - (12 marks) Correct optimal policies
   - (6 marks) Correct descriptions of how the optimal policy changes from one range to the next

3. Determine how the optimal policy changes when $R(s)$ changes from $-0.08$ to $-0.03$.

   You only need to consider the values of $R(s)$ up to two decimal places. That is, you only need to consider $R(s) = -0.08, -0.07, -0.06, -0.05, -0.04, -0.03$.

   **What to submit:**

- Divide up the values of $R(s)$ into ranges such that the optimal policy for each range of values remains the same.
- State each range of values and state the optimal policy. If there is a tie between multiple actions, show all of them as the optimal policy for a state.
- Indicate how the optimal policy changes from the last value in one range to the first value in the next range.

**Marking Scheme:** (24 marks)

- (6 marks) Correct ranges of values
- (12 marks) Correct optimal policies
- (6 marks) Correct descriptions of how the optimal policy changes from one range to the next