# P9120 - Homework # 4

Assigned: November 21st, 2024 Due: 12pm EST on Monday Dec 9th, 2024

**Maximum points that you can score in this Homework is 20.**
**Please include all R or Python code you used to complete this homework.**

1. (10 points) Consider the gridworld depicted below. Each cell, numbered from 1 to 14, allows four possible actions: up, down, left, and right. These actions deterministically move the agent one cell in the corresponding direction unless the move would cause the agent to get off the grid. In such cases, the agent remains in its current location. Executing any action from states 1 to 14 is associated with a reward $R_s$. For the yellow death cell (state 0), taking any action results in a reward of $R_y$ and terminates the episode. Similarly, for the blue target cell (state 15), taking any action provides a reward of $R_b$ and also ends the episode.

   Assume the discount factor $\gamma = 1$, $R_y = -5$ and $R_b = +5$.


| | | | |
|---|---|---|---|
| 0 | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 |

   (a) Suppose $R_s = +1$. Consider an equiprobable random policy $\pi$ (i.e., all actions equally likely). Implement policy evaluation algorithm to compute $V_\pi$. Report the value function on a grid, similar to the example shown on slide 26 of lecture 10. Please include R or Python code.

   (b) Suppose $R_s = +1$, and $\pi$ is an equiprobable random policy. Based on results from part (a), answer the following questions:

   What is $Q_\pi(11, \text{down})$?

   What is the value of $Q_\pi(7, \text{down})$?

   Suppose a new state 16 is added to the gridworld just below state 13, and its actions, left, up, right, and down, take the agent to states 12, 13, 14,and 16, respectively, and yield a reward of +1. Assume that the state transition

probabilities from the original states are unchanged (i.e., action down from state 13 will leave the state unchanged).What is $V_\pi(16)$?

(c) Suppose $R_s$ can take values $-1, 0$ or 1. Which of the three values would cause the optimal policy to return the shortest path to the blue target cell? Derive the optimal policy and optimal value function analytically. Draw the optimal policy on the grid and represent the optimal value function on a separate grid, similar to the example shown on slide 14 of lecture 11.

(d) Note that in part (c), the optimal policy from any of the states 1 to 14 never leads to termination in the yellow cell. Now, suppose $R_s$ can take any value. Provide a value of $R_s$ such that there exist states among 1 to 14 where following the optimal policy results in termination in the yellow cell. Provide a reasoning for your answer.

2. (10 points) In class, we implemented a tabular Q-learning algorithm for the Cart-Pole environment from OpenAI Gym. In this problem, you will modify that implementation by changing some of its parameters based on RL_cartpole_update.ipynb file in Python Lab material folder. These modifications may include, but are not limited to:

   - Adjusting $\epsilon$ in $\epsilon$-greedy policy.
   - Trying different values for the learning rate $\alpha$. You may also incorporate learning rate decay.
   - Modifying the discrete state configuration: in the provided Python code, the four continuous states are discretized into discrete variables with 2, 1, 4, and 3 categories, respectively. You may explore changing the number of categories for each state.
   - Varying the number of training episodes and the maximum number of steps per episode.

   Please do not change the discount factor ($\gamma = 1$) or the number of episodes and the maximum number of steps per episode when evaluating the estimated optimal policy on CartPole.

   Try at least 8 different parameter combinations. For each configuration, detail in writing the settings you used and include the "Performance of Estimated Optimal Policy on CartPole" plot. Provide a discussion analyzing the results of your experiments.