ID6002W: Online Learning and Reinforcement Learning **Release Date**: July 15, 2023

Assignment 2

Roll No: Name:

References (if any):

1. **Programming Question**
   The aim of the exercise is to familiarize you with various learning control algorithms. The goal is to solve several variants of the puddle world problem, shown in Figure 1. This is a typical grid world, with 4 stochastic actions. The actions might result in movement in a direction other than the one intended with a probability of 0.1. For example, if the selected action is N, it will transition to the cell one above your current position with probability 0.9. It will transition to one of the other neighbouring cells with probability 0.1/3. Transitions that take you off the grid will not result in any change.

   There is also a gentle Westerly blowing, that will push you one **additional** cell to the east, regardless of the effect of the action you took, with a probability of 0.5.[1]

   The episodes start in one the start states in the first column, with equal probability. There are three variants of the problem, A, B, and C, in each of which the goal is in the square marked with the respective alphabet. There is a reward of +10 on reaching the goal. There is a puddle in the middle of the gridworld, which the agent likes to avoid. Every transition into a puddle cell, gives a negative reward, depending on the depth of the puddle at that point, as indicated in the figure.

   (a) Implement Q-learning to solve each of the three variants of the problem. For each variant run experiments with a gamma value of 0.9. Turn in two learning curves for each experiment, one that shows how the average number of steps to goal changes over learning trials and the other that shows how the average reward per episode changes. Compute the averages over 50 independent runs. Also indicate the optimal policies arrived at in each of the experiments.

   (b) Repeat part 1 with Sarsa. For both parts, pick a learning rate that seems to best suit the problem.

   Note: Please note that you have to write code for puddle world environment on your own.

2. True or False: If a policy π is greedy with respect to its own value function, then it is optimal. Explain your reasoning.

3. You go to a Halloween party at a mysterious haunted house. Through exploration, you discover that it has the following characteristics. You can either be scared or not scared, and you can either be upstairs or downstairs. If you are scared, running up or down the stairs costs you a unit of energy (reward = -1) but changes your scared state. If you aren't scared, you can run up or down the

[1] For task C you might want to turn off the wind.

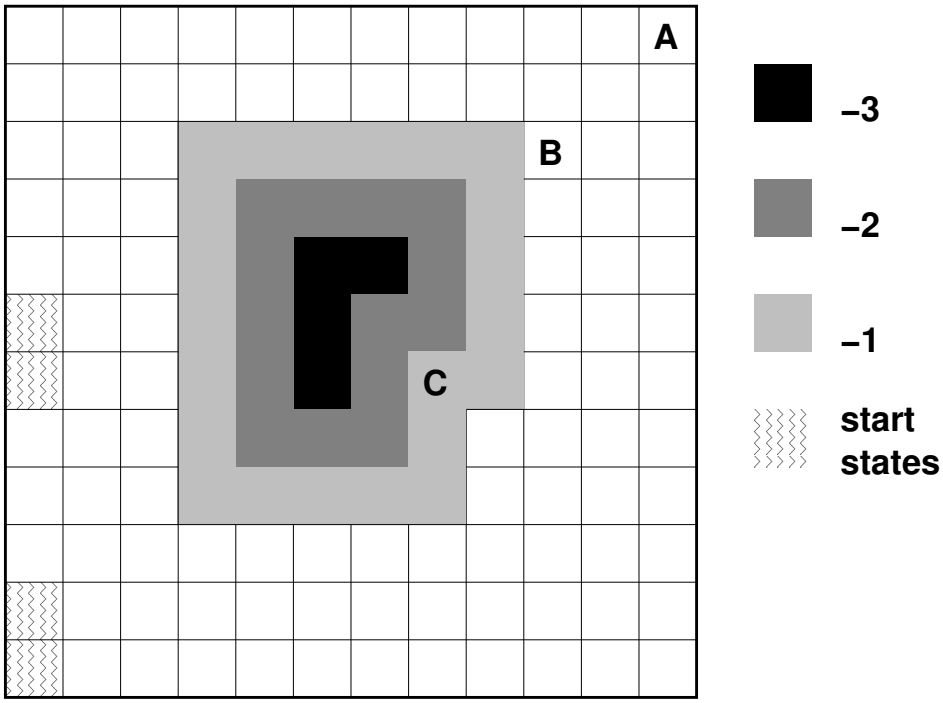

Figure 1: The Puddle World

stairs with a unit energy cost but also a +1 reward for continuing to not be scared (so reward = 0). There are more ghosts downstairs than upstairs, so sitting still in a scared state is worse downstairs (reward = -3) than upstairs (reward = -2). If you sit still while you are not scared, a ghost will pop out and scare you 25% of the time (-2 reward). Otherwise you will remain happy (+2 reward).

(a) Formulate this problem as an MDP (for the sake of uniformity, formulate it as a continuing discounted problem with $\gamma = 0.9$.) The rewards are specified in the description. Explicitly give the state set, action sets, transition probabilities $P^a_{ss'}$ and reward expectations $R^a_{ss'}$. You may do this with a transition graph.

(b) Starting with the policy of running from every state, perform ONE iteration of policy iteration (by hand!). This means evaluate the given policy and then do policy improvement ONCE. Show all steps. What is the resulting policy?

4. Briefly explain the ideas behind the success of DQN.

5. For the case in which the return is expected sum of discounted future rewards and π is a fixed deterministic policy, give the Bellman equation for V^π and V^*.