In this assignment, you are required to implement the grid world domain with minor modifications (the transition probabilities are different from the one in lecture). Recall that the goal of the agent is to start from the lower corner square (say (0,0)) and go to the top right corner square (3,4). There are 2 unreachable squares (the shaded ones). The agent receives a reward of +10 when it reaches the goal square, −50 when it falls into the pit and −1 for every action that it takes (i.e., −1 for every time-step).

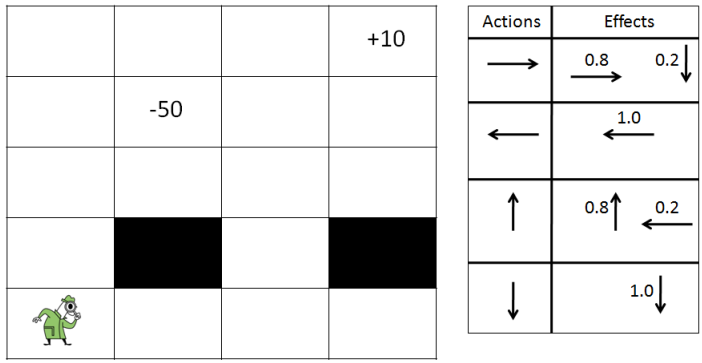

Figure 1: Grid world domain and the effects of action.

In this assignment you are required to implement, value iteration and Q-learning. Below are the required outputs. Please note that for the grid values, you can hand draw the grid and write the values.

1. Present the learned values using value iteration. The output must be similar to the one on the class slides (slide 37 with the values in every square).

2. Now assume that the transition probabilities and rewards are not known in advance. We will use Q-learning (model-free) RL to learn in this setting using an epsilon-greedy policy with $\epsilon = 0.5$ i.e., with 0.5 probability choose the greedy action and with 0.5 choose a random action uniformly. Decrease the exploration every 10 iterations. Present the Q-values learned for each of the action in each state (Q(s,a)) in the grid. A sample output is presented in Figure 2.

3. Finally, make the rewards at each time-step to be 0 and repeat the experiments. Report the new results.

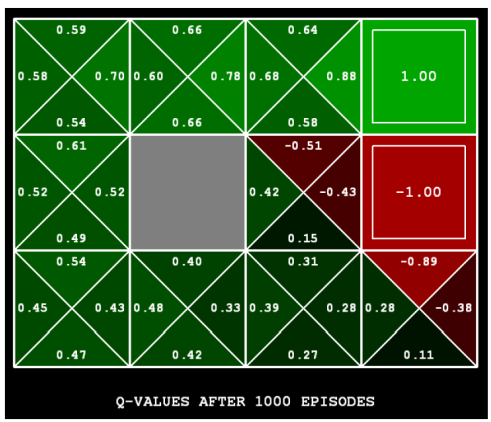

Figure 2: An Example Q-value in a different grid world domain. Note that the values are presented for all actions in all the states.