Reinforcement Learning Tutorial Exercises

Exercise 1

Consider the ϵ-greedy action selection in reinforcement learning problems, in the case of two actions and $\epsilon = 0.5$. What is the probability that the greedy action is selected?

a) 0.5

b) 0.25

c) 0.75

d) 0

e) 1.0

Exercise 2

Consider the gridworld problem shown in Figure 1. An agent needs to move to the target goal position **G** starting from any cell (arbitrary position) by following the optimum policy which gathers the largest reward.

The immediate rewards $r(s, a)$ for the transition from one state to another for this problem are also shown in Figure 1. For example, the reward received by moving from state s_1 to s_2 by taking the *right* action is 0. The reward for moving from state s_5 to state G (i.e. taking the *north* action) is 100.

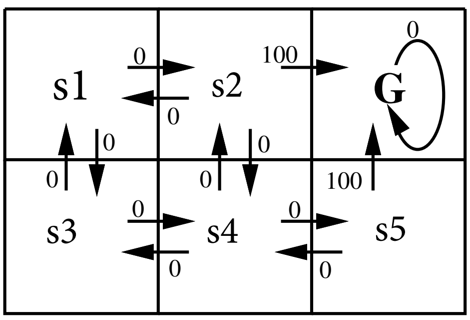

Figure 1: $r(s, a)$ (immediate reward) values.

Now, consider the application of the Q learning algorithm in the above reinforcement learning gridworld problem. Reminder: the Q values are calculated using the following formula:

$$\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s',a')$$

The current values of $\hat{Q}$ (i.e. in the current iteration) are shown in Figure 2. For example, $\hat{Q}(s_1, a_{right}) = 72$, $Q(s_2, a_{left}) = 63$, $\hat{Q}(s_3, a_{north}) = 63$.

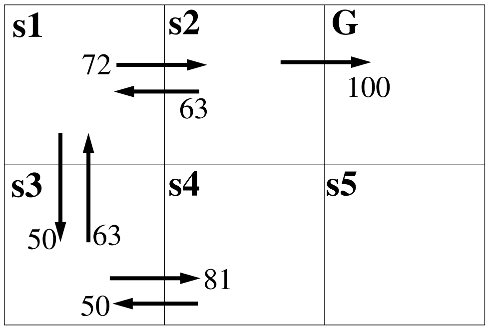

Figure 2: Current values of $\hat{Q}(s,a)$ for all state-action pairs.

What will be the new value for $\hat{Q}(s_1, a_{down})$ after applying the above Q-learning formula for one iteration, if $\gamma = 0.9$?

a) 63

b) 81

c) $0.9 \times 72 = 64.8$

d) $0.9 \times 63 = 56.7$

e) $0.9 \times 81 = 72.9$

f) 90

Exercise 3

Implement in a programming language of your choice (Python or Java) the Q-learning applied to the problem of the previous Exercise 2, until the Q values converge and they do not change any more.

Hint: You need to consider multiple episodes, i.e. after the terminal (final, goal state) is reached, a new episode is started. The initial values of the Q values for the new episode will be the ones that the previous episode had calculated.

Check your derived Q values with the Figure 4 in the lecture slide 14.

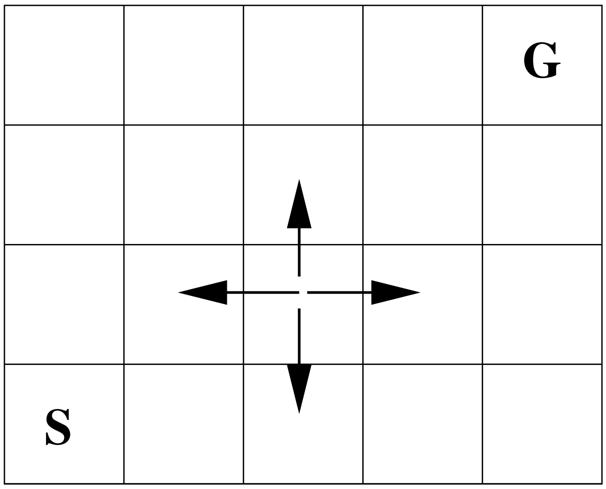

Figure 3: The grid world example for the application of reinforcement learning techniques.

Exercise 4

Consider the grid in Figure 3. An agent can move in either of the four directions starting from S and finishing in the goal state G.

1. If the reward on reaching the goal is 100, and all other rewards between state transitions are 0, write a program in a programming language of your choice (e.g. Java, Python, C++) which uses Q learning to learn the optimal policy. Assume that $\gamma = 0.9$.
2. What are the actions of the optimal policy?

Exercise 5

In Exercise 4, how does the optimal policy change if another goal state is added to the lower right corner with reward -100.

Hint: To see this, rerun your implementation with the additional goal.