CS6700 : Reinforcement Learning

Homework #4 **Deadline: 18th October 2018, 5pm**

Instructions:

- Submit well-documented code on moodle and printout of report after class.
- Any kind of plagiarism will be dealt with severely. Acknowledge every resource used.

Question 1

1. The objective of this question is to compare value iteration and policy iteration on a 10×10 gridworld based on the actions, rewards and the state space given below.

   - **State space**: Gridworld has 100 distinct states. There are two variants of this gridworld, one with a terminal state as Goal 1 and other with Goal 2. For the variant with Goal 1 as a terminal state, Goal 2 is treated as a normal state and vice-versa. There are two wormholes labeled as IN in Grey and Brown, any action taken in those states will teleport you to state labeled OUT in Grey and Brown respectively. In case of Grey wormhole you can teleport to any one of the states labelled OUT with equal probability (i.e. 1/4). States labeled OUT is just a normal state. An instance of this gridworld is shown in the figure below.

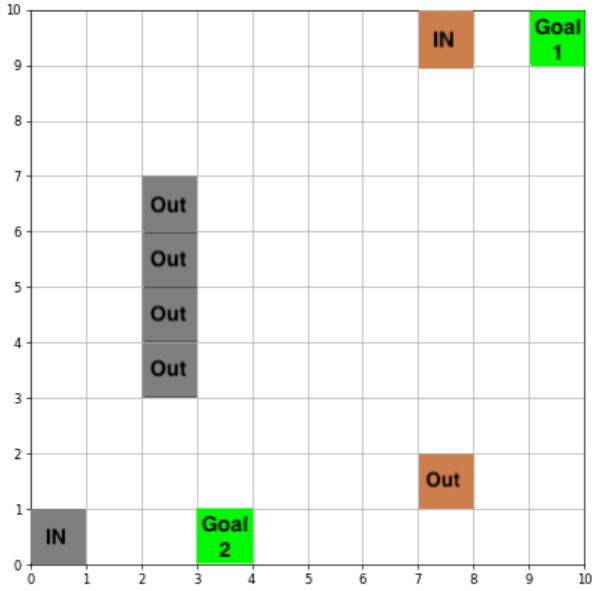

Figure 1: 10×10 gridworld

   - **Actions:** In each non-terminal state, you can take 4 actions $\mathcal{A} = \{\text{Up, Down, Left, Right}\}$, which moves you one cell in the respective direction.

- **Transition model:** Gridworld is stochastic because the actions can be unreliable. In this model, action "X" (X can be Up, Down, Left, or Right) moves you one cell in the X direction of your current position with probability 0.8, but with probabilities 0.2/3 it will transition to one of the other neighbouring cells. Transitions that take you off the grid will not result in any change.
- **Rewards:** The reward is 0 on all transitions until the terminal state is reached. Reaching terminal state gives you a reward of +10. Take the discount factor $\alpha = 0.7$.

1. Implement value iteration and policy iteration. Start with $J_0(s) = 0$ and $\pi_0(s) =$ UP, $\forall s$. (3 marks)

2. Answer the following questions for variant with Goal 1 as terminal state: (2+2+2+2+2 marks)

   (a) Plot graph of $\max_s |J_{i+1}(s) - J_i(s)|$ vs iterations and $\sum_s \pi_{i+1}(s) \neq \pi_i(s)$ vs iterations for both value iteration and policy iteration.

   (b) Compare value iteration and policy iteration by plotting $J(s)$ vs iterations for three random states. Which converges faster? Why?

   (c) Show $J(s)$ and greedy policy $\pi(s)$, $\forall s$, obtained after 5 iterations, and after you stop value iteration.

   (d) Show $J(s)$ and policy $\pi(s)$, $\forall s$, obtained after 5 iterations, and after you stop policy iteration.

   (e) Explain the behaviour of J and greedy policy π obtained by value iteration and policy iteration.

3. Answer the following question for variant with Goal 2 as terminal state: (2 marks)

   (a) Show $J(s)$ and policy $\pi(s)$, $\forall s$, obtained after you stop value iteration and policy iteration and explain it's behaviour.

Note: You have to show $J(s)$ and policy $\pi(s)$ with arrows, $\forall s$, on the image of girdworld, not in some form of table unless you explicitally make your table look like gridworld highlighting wormholes and terminal state.

Question 2

Consider a problem of a taxi driver, who serves three cities A, B and C. The taxi driver can find a new ride by choosing one of the following actions.

1. Cruise the streets looking for a passenger.

2. Go to the nearest taxi stand and wait in line.

3. Wait for a call from the dispatcher (this is not possible in town B because of poor reception).

For a given town and a given action, there is a probability that the next trip will go to each of the towns A, B and C and a corresponding reward in monetary units associated with each such trip. This reward represents the income from the trip after all necessary expenses have been deducted. Please refer Table 1 below for the rewards and transition probabilities. In Table 1 below, p_{ij}^k is the probability of getting a ride to town j, by choosing an action k while the driver was in town i and r_{ij}^k is the immediate reward of getting a ride to town j, by choosing an action k while the driver was in town i.

Town i	**Actions** k	**Probabilities** p_{ij}^k	**Rewards** r_{ij}^k
		j = A B C	j = A B C
A	1 2 3	A B C $\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/16 & 3/4 & 3/16 \\ 1/4 & 1/8 & 5/8 \end{bmatrix}$	A B C $\begin{bmatrix} 10 & 4 & 8 \\ 8 & 2 & 4 \\ 4 & 6 & 4 \end{bmatrix}$
B	1 2	A B C $\begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/16 & 7/8 & 1/16 \end{bmatrix}$	A B C $\begin{bmatrix} 14 & 0 & 18 \\ 8 & 16 & 8 \end{bmatrix}$
C	1 2 3	A B C $\begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/8 & 3/4 & 1/8 \\ 3/4 & 1/16 & 3/16 \end{bmatrix}$	A B C $\begin{bmatrix} 10 & 2 & 8 \\ 6 & 4 & 2 \\ 4 & 0 & 8 \end{bmatrix}$

Table 1: Taxi Problem: Probabilities and Rewards

Suppose $1 - \beta$ is the probability that the taxi will breakdown before the next trip. The driver's goal is to maximize the total reward untill his taxi breakdown.

Implement the following algorithms to solve the taxi problem.

1. Find an optimal policy using **policy iteration** starting with a policy that will always cruise independent of the town. Solve it for discount factors β ranging from 0 to 0.95 with intervals of 0.05. Tabulate the optimal policies and optimal values obtained for different values of β. (5 marks)

2. (a) Find an optimal policy using **modified policy iteration**. Let $m_k = 5\ \forall k$. Start with a policy that will always cruise independent of the town. Let $\beta = 0.9$. What are the optimal values? (3 marks)

   (b) Do you find any improvement if you choose $m_k = 10\ \forall k$? Explain. (2 marks)

3. Find an optimal policy using **value iteration** and **Gauss-Seidel value iteration** starting with a zero vector. Let $\beta = 0.9$. What are the optimal values? (5 marks)