CS6700 : Reinforcement Learning

Homework #2 **Deadline: 24th September 2018**

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

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.

Table 1: Taxi Problem: Probabilities and Rewards

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

Suppose that after N rides the driver will stop working for that day and his goal is to maximize the total reward he accumulate for that day.

1. Implement the DP algorithm to solve this problem for N=10 and N=20. (4 Marks)
2. Find out the optimal policy for N=10 and N=20. (5 Marks)
3. Consider a policy that always forces the driver to go to the nearest taxi stand, irrespective of the state. Is it optimal? Justify your answer. (3 Marks)

Question 2

1. The objective of this question is to implement value 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. States labeled OUT is just a normal state. An instance of this gridworld is shown in the figure below.

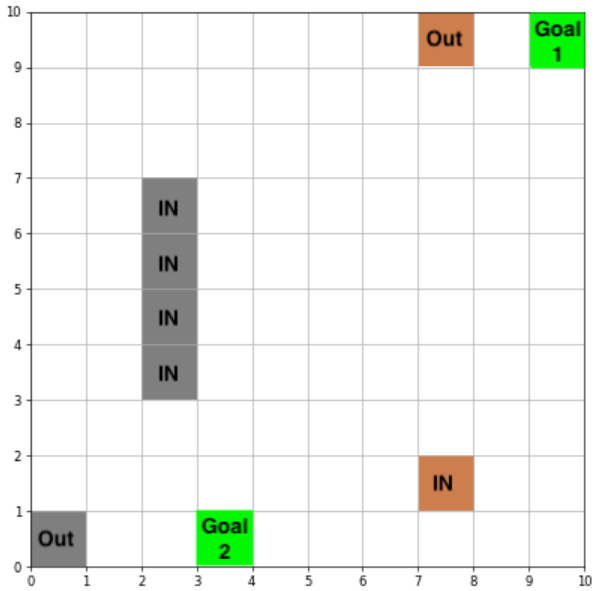

Figure 1: 10×10 gridworld

   - **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 your current position with probability 0.8, but with probabilities 0.1 and 0.1 moves you one cell at angels of $90°$ and $-90°$ to the direction X, respectively. For example, if the selected action is UP, it will transition to the cell one above your current position with probability 0.8, one cell to the left with probability 0.1 and one cell to the right with probability 0.1. Transitions that take you off the grid will not result in any change.

- **Rewards:** The reward is -1 on all transitions until the terminal state is reached. Reaching terminal state gives you a reward of $+100$.

Implement value iteration, which has the following pseudocode: (6 marks)

Initialize $J_0(s) = 0, \forall s$

For $i = 0, 1, 2, \ldots$

- $J_{i+1}(s) = \max_a \sum_{s'} P_{ss'}(a)[r(s, a, s') + J_i(s')], \forall s$

We additionally define the greedy policy w.r.t J_i as,

$$\pi_i(s) = \operatorname*{argmax}_a \sum_{s'} P_{ss'}(a)[r(s, a, s') + J_i(s')], \forall s$$

Answer the following questions for both variants of the gridworld: (1+2+6+3 marks)

1. From the pseudocode give above, observe that for loop goes on till infinity, so when would you decide to stop value iteration?
2. Plot graph of $\max_s |J_{i+1}(s) - J_i(s)|$ vs iterations.
3. Show $J(s)$ and greedy policy $\pi(s)$, $\forall s$, after 10 iterations, 25 iterations, and after you decide to stop value iteration.
4. Explain the behaviour of J and greedy policy π obtained after you decide to stop value iteration.