EECE 5698 - ST: Reinforcement Learning Spring 2023

HW2

Problem 1.

Consider the following system with two states $s_k \in \{s^1 = 0, s^2 = 1\}$.
There are two possible actions: a^1 and a^2. The transition probabilities can be expressed as:

$$p(s'|s,a^1)\begin{cases}1 & s=0, s'=0\\ 0 & s=0, s'=1\\ 0 & s=1, s'=0\\ 1 & s=1, s'=1\end{cases} \qquad p(s'|s,a^2)\begin{cases}0 & s=0, s'=0\\ 1 & s=0, s'=1\\ 1 & s=1, s'=0\\ 0 & s=1, s'=1\end{cases}$$

Reward function is as follows: $\begin{cases} moving\ to\ state\ s^2: +1 \\ moving\ to\ state\ s^1: 0 \\ action\ a^1\ and\ a^2: 0\end{cases}$

Start with a random policy $\pi^0(s^1) = a^1, \pi^0(s^2) = a^1, \gamma = 0.9, \theta = 0.85$. Use Policy Iteration to compute $\pi^1(s^1)$, $\pi^1(s^2)$. Use $V_0(s^1) = V_0(s^2) = 0$, for initialization of Policy Evaluation.

Problem 2.

Consider the problem defined in Problem 1.

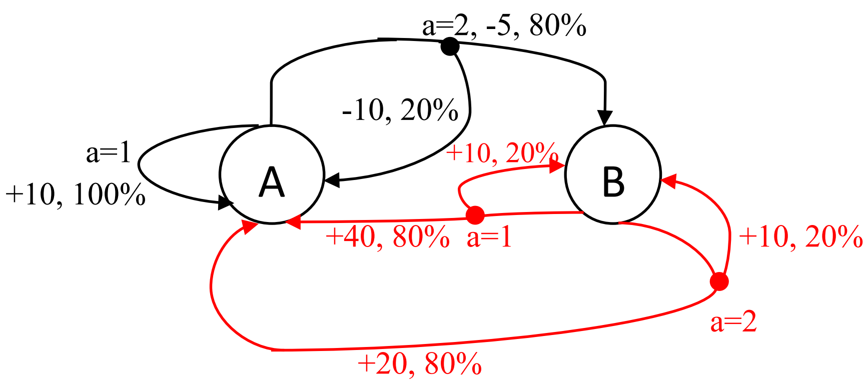
a) Given $\begin{bmatrix} V_0(s^1) \\ V_0(s^2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\gamma = 0.9$, perform Value Iteration method to compute V_1, V_2, V_3 .

b) Compute $\pi(s = 0)$ and $\pi(s = 1)$ associated with V_3.

Problem 3.

Consider the following MDP having two states: A, B. In each state, there are two possible actions: 1 and 2. The transition model and reward are shown in the diagram below.
Apply Policy Iteration to determine the optimal policy and state values of A and B.
Assume the initial policy is action 2 for both staters, $\gamma = 0.9$.
For evaluation of policy, you need to solve two set of linear equations for the following form, instead of iterative steps of policy evaluation:

$$V^{\pi}(s) = \sum_{s',r} P(s'|s,\pi(s))[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$



*Here is an example of transition and reward from the diagram:
In state A, action 2 moves the agent to state B with probability 0.8 with the corresponding reward -5, and make the agent stay at state A with probability 0.2 and corresponding reward -10.

Problem 4.

Consider the following maze with 14 states and a goal. The agent can take one of the following four actions at any given state $A = \{UP, Down, Right, Left\}$

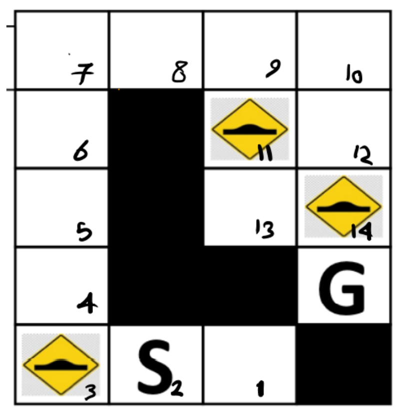
The state transitions are deterministic; for example $P(s'=10 \mid s=12, a=U)=1$

The reward is as follows: $\begin{cases} -1 & \text{taking any action} \\ +200 & \text{moving to goal} \\ -10 & \text{moving to bump} \end{cases}$

a) Using $\gamma=1$ and $\theta=0.5$, perform Vector-form Value Iteration method with $V_0(s)=0$ to compute V^*.

b) Compute the optimal Policy.

* Show all intermediate state values in maze, without details of calculation.



Problem 5.

For an MDP defined by state space S, action space A, Reward $R(s,a,s')$ and transition probability $p(s'|s,a)$, write the following:

a) For a given policy π, write

- $V^\pi(s)$ based on V^π
- $V^\pi(s)$ based on Q^π
- $Q^\pi(s,a)$ based on V^π
- $Q^\pi(s,a)$ based on Q^π

b) For the optimal policy π^*, write

- $V^*(s)$ based on V^*
- $V^*(s)$ based on Q^*
- $Q^*(s,a)$ based on V^*
- $Q^*(s,a)$ based on Q^*

- Help: An example of response:

$$v^\pi(s) = \sum_{s'} p\big(s' \big| s, \pi(s)\big)[R(s,\pi(s),s') + \gamma V^\pi(s')]$$

Questions about the HW should be directed to TA, Begum Taskazan, at taskazan.b@northeastern.edu.