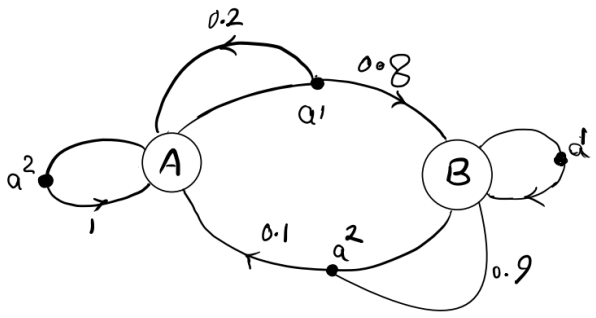
Problem 1.

Consider the following system with the state space $S=\{A,B\}$, and action space $A=\{a^1, a^2\}$. The state transition diagram is shown below, where $P(s'=B \mid s=A, a=a^1)=0.8$, $P(s'=A \mid s=A, a=a^1)=0.2$.

The reward is as follows:

$$\begin{cases} +2 & \text{moving to state } B \\ 0 & \text{moving to state } A \\ -1.5 & \text{taking action } a^1 \\ -1 & \text{taking action } a^2 \end{cases}$$



a) Construct transition matrices $M(a^1)$, $M(a^2)$ and compute $R_s^{a^1}$, $R_s^{a^2}$.

b) Perform matrix-form policy Iteration method with initial policy $\pi^1(A)=a^2$, $\pi^1(B)=a^1$ and $\gamma=0.9$ to compute π^*.

Problem 2.

For the system defined in Problem 2, perform matrix-form Value Iteration method with $V_0(S)=0$, $\gamma=0.9$ and $\theta=0.5$ to compute V^* and π^*.

Problem 3.

Consider an MDP with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transitions are governed through the following transition matrices:

$$M(a^1)=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M(a^2)=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The reward is as follows:

$$\begin{cases} +5 & \text{moving to State } B \\ 0 & \text{moving to State } A \\ -1 & \text{taking action } a^2 \\ 0 & \text{taking action } a^1 \end{cases}$$

Consider an initial policy $\pi^0=\begin{bmatrix} \pi^0(A) \\ \pi^0(B) \end{bmatrix}=\begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$, $\gamma=0.9$ and episode length 5. Perform Monte Carlo Policy Iteration method to obtain the best policy.

*you need to show all trajectories, the approximation of Q-values and Policy Improvement till the time that Policies in two Consecutive Iterations stays the same.

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