Problem 1.

Consider the following system with two states $\{A, B\}$ and two possible actions $\{a^1, a^2\}$.

The transition probabilities can be expressed as:

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

The reward function is as follow: $\begin{cases} 2 & \text{moving to state } s'=B \\ 0 & \text{moving to state } s'=A \\ -1 & \text{taking action } a^2 \\ 0 & \text{taking action } a^1 \end{cases}$

a) Consider the initial Q-value $\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix}$ and $\begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$ with $\gamma=0.9$ and $\alpha=0.5$. Perform Q-Learning Algorithm with $\varepsilon=0$ (greedy policy), initial state $s_0=A$ for four steps: $(s_0=A, a_0), (s_1, a_1), (s_2, a_2), (s_3, a_3), (s_4, a_4)$. Show all intermediate Q-Values.

b) Compute the final policy $\pi(A)$, and $\pi(B)$, after all transitions in part (a).

c) If you would use SARSA istead of Q-Learning, would the intermediate Q-Values and final policy be different from part (a). Justify your answer. (no computations needed in this part).

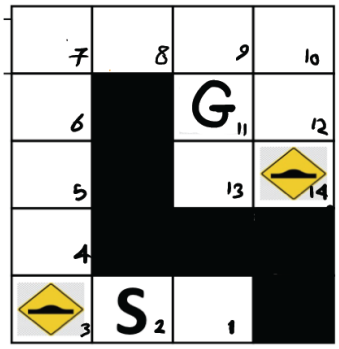
Problem 2.

Consider the following maze problem with 14 states and four actions $A=\{U, R, D, L\}$. The state transitions are deterministic, and reward for taking any action is -1 and moving to the goal state 100 and bump -10.

a) Set all initial Q-values to zero; $Q(s,a)=0$ for all $s \in S$, $a \in A$, $\alpha=0.5$ and $\gamma=0.9$. Run one episode of Q-Learning Algorithm when agent starts from state 5 and follows greedy policy. Note that in the case of equal Q-values, the tie break for actions will be in the following order U, R, D, L. For example

$$\underset{a\in\{U,R,D,L\}}{\operatorname{argmax}} Q(s,a) = \begin{cases} U & \text{if } Q(s,U)=Q(s,R)=Q(s,D)=Q(s,L) \\ R & \text{if } Q(s,R)=Q(s,D)=Q(s,L) > Q(s,U) \\ D & \text{if } Q(s,L)=Q(s,D) > Q(s,U), Q(s,R) \end{cases}$$

b) Show step by step transitions, Q-values and final policy obtained in this single episode.



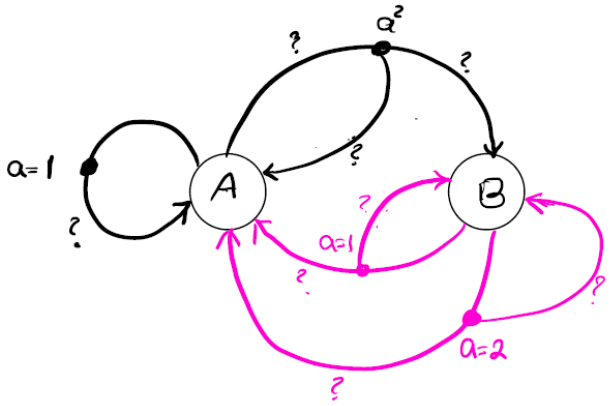
Problem 3.

Consider the following system with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transition is unknown and learning should be achieved through interactions. Consider the following state-action-reward obtained through Softmax policy in Actor-Critic algorithm.

$(S_0 = A, a_0 = a^1, r = 10), (S_1 = A, a_1 = a^2, r = -5), (S_2 = B, a_2 = a^1, r = 40),$

$(S_3 = A, a_3 = a^2, r = -5), (S_4 = B, a_4 = a^2, r = 20), (S_4 = A, a_4 = a^1, r = +10), S_5 = A$

Set the initial preferences and state values to zero. Use $\alpha = 0.5$, $\beta = 0.1$ and $\gamma = 0.9$ and show all intermediate preferences, state values and policies.



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