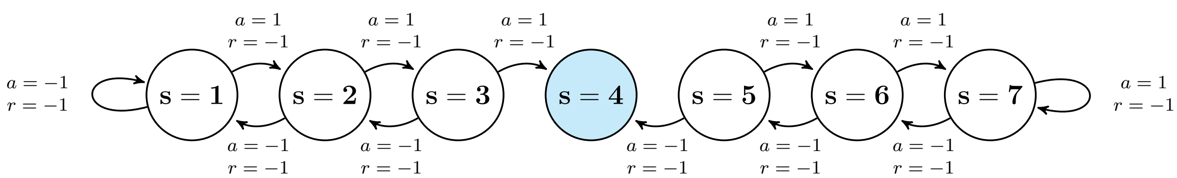
IE-456/556 & EEE-448/548
Reinforcement Learning and Dynamic Programming
Homework Assignment 4 – Due July 27 23:59

We have an assembly line which is modeled as an MDP with 7 states and 2 actions of $a \in \{-1, 1\}$ where the discount is $\gamma = 1$. We consider deterministic transitions for this MDP such that $s' = s + a$ with two exceptions: by taking $a = -1$ at $s = 1$ we stay in $s = 1$, and by taking $a = 1$ at $s = 7$ we stay in $s = 7$. We also have the terminal state $s = 4$ as our goal and taking any action from it ends the episode with a reward of $r = 0$. However, from all other states, any action incurs a reward of $r = -1$.



By inspection, we can see that $v^*(s) = -|s - 4|$.

(a) (**30 pts**) We intend to perform tabular Q-learning on this MDP with the learning rate $\alpha = 0.5$ and initial values of zero for all Q values. Suppose we observe the following trajectory in the form of *(state, action, reward)*:

$$(3, -1, -1),\ (2, 1, -1),\ (3, 1, -1),\ (4, 1, 0)$$

Use the tabular Q-learning to find updated values for

$$Q(3, -1),\ Q(2, 1),\ Q(3, 1)$$

In addition, we intend to perform linear function approximation together with Q-learning. To this end, we consider three features with the following entry for each state-action pair:

$$\begin{bmatrix} s \\ a \\ 1 \end{bmatrix}$$

Given the weights **w** and a single tuple (s, a, r, s'), the loss function will be

$$J(\mathbf{w}) = \Big(r + \gamma \max_{a'} \hat{Q}(s', a', \mathbf{w}^-) - \hat{Q}(s, a, \mathbf{w})\Big)^2$$

where $\hat{Q}(s', a', \mathbf{w}^-)$ is a target network parametrized by fixed weights $\mathbf{w}^-$.

(b) (**35 pts**) Find an expression for the gradient $\nabla_{\mathbf{w}} J(\mathbf{w})$ and write out what is the formula for updating the weights to get new weights w'.

(c) (**35 pts**) Suppose we currently have weight vectors

$$\mathbf{w} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \ \mathbf{w}^{-} = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}$$

and we observe a tuple $(s = 2, a = -1, r = -1, s' = 1)$. Perform a single gradient update to the parameters $\mathbf{w}$ given this sample with the learning rate $\alpha = 0.25$.