# CS234 Problem Session

Week 3: Jan 27

## 1) [Breakout Rooms] Q-learning Practice

Consider an unknown MDP with three states $(A, B, C)$ and two actions $(\leftarrow, \rightarrow)$. Suppose the agent chooses actions according to some policy $\pi$ in the unknown MDP, collecting a dataset consisting of samples $(s, a, s', r)$ representing taking action $a$ in state $s$ resulting in a transition to state $s'$ and a reward of $r$.

| $s$ | $a$ | $s'$ | $r$ |
|---|---|---|---|
| $A$ | $\rightarrow$ | $B$ | 2 |
| $C$ | $\leftarrow$ | $B$ | 2 |
| $B$ | $\rightarrow$ | $C$ | $-2$ |
| $A$ | $\rightarrow$ | $B$ | 4 |

You may assume a discount factor of $\gamma = 1$.

Recall the update function of Q-learning is:

$$Q(s_t, a_t) = (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot (r_t + \gamma max_{a'} Q(s_{t+1}, a')) \tag{1}$$

Assume that all Q-values are initialized to 0, and use a learning rate of $\alpha = \frac{1}{2}$.

(a) Run Q-learning on the above experience table and fill in the following Q-values:

$Q(A, \rightarrow) =?$

$Q(B, \rightarrow) =?$

(b) After running Q-learning and producing the above Q-values, you construct a policy $\pi_Q$ that maximizes the Q-value in a given state: $\pi_Q(s) = argmax_a Q(s, a)$.

What are the actions chosen by the policy in states $A$ and $B$?

(c) Compute the MLE MDP model estimates of the transition function $\hat{P}(s, a, s')$ and reward function $\hat{R}(s, a, s')$.

Write down the following quantities. You may write N/A for undefined quantities.

$\hat{P}(A, \rightarrow, B) =?$

$\hat{P}(B, \rightarrow, A) =?$

$\hat{P}(B, \leftarrow, A) =?$

$\hat{R}(A, \rightarrow, B) =?$

$\hat{R}(B, \rightarrow, A) =?$

$\hat{R}(B, \leftarrow, A) =?$

## 2) [Breakout Rooms] Value Functions

Prove that the following two definitions of the state-value function are equivalent:

$$V^{\pi}(s) = \mathbf{E}[G_t | S_t = s, \pi] \tag{2}$$

$$V^{\pi}(s) = \mathbf{E}[G | S_0 = s, \pi] \tag{3}$$

.

## 3) [Breakout Rooms] Negative Reward MDP

Consider a finite MDP with bounded rewards, where all rewards are negative. That is, $R_t < 0$ always. Let $\gamma = 1$. The MDP is finite horizon, with horizon $L$, and also has a deterministic transition function and initial state distribution (rewards may be stochastic). Let $H_\infty = (S_0, A_0, R_0, S_1, A_1, R_1, ...S_{L-1}, A_{L-1}, R_{L-1})$ be any history that can be generated by a deterministic policy *pi*. Prove that the sequence $V^\pi(S_0), V^\pi(S_1), ...V^\pi(S_{L-1})$ is strictly increasing.