CS234 Problem Session

Week 5: Feb 10

1) [CA Session] Mars Rover REINFORCE

s_1	s_2	s_3	s_4	s_5	s_6	s_7
+1	+0	-1		-1	+0	+10

Figure 1: Mars Rover MDP

Let us consider the Mars Rover MDP seen in Figure 1. Similar to the in class example, s_1 and s_7 are terminal states. The rewards are received when you enter a state (the reward for entering state s_4 is 0). There are two actions, TryLeft and TryRight. TryLeft transitions from state s_i to s_{i-1} with 0.5 probability and stays in state s_i with 0.5 probability. Similarly, TryRight transitions from state s_i to s_{i+1} with 0.5 probability and stays in state s_i with 0.5 probability. Let $\gamma = 1$.

We want to apply REINFORCE to learn a policy in this Mars Rover setting. Let our feature representation be a one-hot encoding using the state, action pair. More concretely, let us denote $a_1 =$ TryLeft and $a_2 =$ TryRight. Then our feature representation is $\phi(s_i, a_j)_k = 1$ if $((j-1)*7) + (i-1) = k$ and 0 otherwise (assuming the vector is 0-indexed). Let us use a softmax policy parameterized by θ:

$$\pi_\theta(s,a) = e^{\phi(s,a)^T\theta} / \sum_a e^{\phi(s,a)^T\theta}$$

(a) What is the score function for this softmax policy?

(b) Using REINFORCE, what is the update equation for θ?

(c) Now let us run the REINFORCE algorithm. Assume θ is initialized to be all zeros. We execute one rollout of the policy π_θ to obtain the following episode:

$$(s_4, a_0, -1, s_3, a_1, 0, s_4, a_1, -1, s_5, a_1, 0, s_6, a_0, 0, s_6, a_1, 10)$$

Run REINFORCE to update θ three times using the provided episode. For simplicity, let $\alpha = 1$.

2) [Breakout Rooms] Gaussian Policy Gradients

Suppose you have a Gaussian policy that samples actions a from a normal distribution with mean $\phi(s)^T\theta$ and variance σ^2.

As a reminder, the Gaussian PDF is as follows:

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

(a) What is $\nabla_\theta \text{log}(\pi(s, a; \theta))$?

(b) What is $\nabla_\sigma \text{log}(\pi(s, a; \theta))$?

3) [Breakout Rooms] Bayes Expressions

Write an expression for the probability that the state at time 0 is s given that the state at time 1 is s' and the action at time 0 is a. Let us define $d_0(s) = Pr(S_0 = s)$. Please write your answer in terms of d, π, and the transition probabilities $P(s, a, s')$. Recall Bayes' Theorem:

$$Pr(A = a|B = b) = \frac{Pr(B = b, A = a}{Pr(B = b)} \tag{1}$$

.