CS234 Problem Session Solutions

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?

Solution The score function is $\nabla_\theta \log \pi_\theta(s,a) = \phi(s,a) - \mathbb{E}_{\pi_\theta}[\phi(s,\cdot)]$

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

Solution $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t = \theta + \alpha[\phi(s,a) - \sum_b \pi_\theta(s,b) \cdot \phi(s,b)] G_t$

(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:

$$(s_4, a_1, -1, s_3, a_2, 0, s_4, a_2, -1, s_5, a_2, 0, s_6, a_1, 0, s_6, a_2, 10)$$

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

Solution After the first update:

$\theta = [0,0,0,0,0,0,0,0,0,0,0,0,0,0] + 1 \cdot [[0,0,0,1,0,0,0,0,0,0,0,0,0,0] - (0.5 \cdot [0,0,0,1,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,0,1,0,0,0])]8$
$= [0,0,0,4,0,0,0,0,0,0,-4,0,0,0]$

After the second update:

$\theta = [0,0,0,4,0,0,0,0,0,0,-4,0,0,0] + 1 \cdot [[0,0,0,0,0,0,0,0,0,1,0,0,0,0] - (0.5 \cdot [0,0,1,0,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,1,0,0,0,0])]9$
$= [0,0,-4.5,4,0,0,0,0,0,4.5,-4,0,0,0]$

After the third update:

$\theta = [0,0,-4.5,4,0,0,0,0,0,4.5,-4,0,0,0] + 1 \cdot [[0,0,0,0,0,0,0,0,0,0,1,0,0,0] - (0.5 \cdot [0,0,0,1,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,0,1,0,0,0])]9$
$= [0,0,-4.5,-0.5,0,0,0,0,0,4.5,0.5,0,0,0]$

Note that instead of updating θ in place, we use the original θ used to collect the data in the computation of π_θ.

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 $\nabla_\theta \log(\pi(s, a; \theta))$?

Solution

$$\begin{aligned}
\nabla_\theta \log(\pi(s, a; \theta)) &= \frac{1}{\pi(s, a; \theta)}\nabla_\theta \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} \\
&= \frac{1}{\pi(s, a; \theta)}\pi(s, a; \theta)\nabla_\theta \frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2 \\
&= \frac{-1}{2\sigma^2}2(a-\phi(s)^T\theta)(-\phi(s)) \\
&= \frac{1}{\sigma^2}(a-\phi(s)^T\theta)(\phi(s))
\end{aligned}$$

Or write down the log density

$$\begin{aligned}
\log \pi(s, a; \theta) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{\phi(s)^T\theta - a}{\sigma}\right)^2 \\
&= -\frac{1}{2}\left(\log(2\pi) + 2\log\sigma + \left(\frac{\phi(s)^T\theta - a}{\sigma}\right)^2\right)
\end{aligned}$$

and differentiate w.r.t. θ:

$$\begin{aligned}
\nabla_\theta \log \pi(s, a; \theta) &= -\frac{1}{2}\nabla_\theta\left(\frac{\phi(s)^T\theta - a}{\sigma}\right)^2 \\
&= -\frac{1}{2}\cdot 2\left(\left(\frac{\phi(s)^T\theta - a}{\sigma}\right)\frac{\phi(s)}{\sigma}\right) \\
&= \frac{a-\phi(s)^T\theta}{\sigma^2}\phi(s)
\end{aligned}$$

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

Solution

$$
\begin{aligned}
\nabla_\sigma \log(\pi(s,a;\theta)) &= \frac{1}{\pi(s,a;\theta)} \nabla_\sigma \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} \\
&= \frac{1}{\pi(s,a;\theta)} [\frac{1}{\sqrt{2\pi\sigma^2}} \nabla_\sigma e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} + e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} \nabla_\sigma \frac{1}{\sqrt{2\pi\sigma^2}}] \\
&= \nabla_\sigma \frac{-1}{2\sigma^2}(a-\phi(s)^T\theta)^2 + \frac{1}{\pi(s,a;\theta)} e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} \frac{1}{\sqrt{2\pi}} \nabla_\sigma \frac{1}{\sigma} \\
&= \frac{1}{\sigma^3}(a-\phi(s)^T\theta)^2 + \frac{1}{\pi(s,a;\theta)} e^{\frac{-1}{2}(\frac{a-\phi(s)^T\theta}{\sigma})^2} \frac{1}{\sqrt{2\pi}} \frac{-1}{\sigma^2} \\
&= \frac{1}{\sigma^3}(a-\phi(s)^T\theta)^2 + \frac{-1}{\sigma}
\end{aligned}
$$

Or, directly differentiating the log density w.r.t. σ,

$$
\begin{aligned}
\frac{\partial}{\partial\sigma} \log \pi(s,a;\theta) &= -\frac{1}{2}\left(2\frac{\partial \log\sigma}{\partial\sigma} + \frac{\partial}{\partial\sigma}\left(\frac{\phi(s)^T\theta - a}{\sigma}\right)^2\right) \\
&= -\frac{1}{2}\left(\frac{2}{\sigma} - \frac{2(\phi(s)^T\theta - a)^2}{\sigma^3}\right) \\
&= \frac{(\phi(s)^T\theta - a)^2}{\sigma^3} - \frac{1}{\sigma}
\end{aligned}
$$

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:

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

Solution

$$\begin{aligned}
Pr(S_0 = s|A_0 = a, S_1 = s') &= \frac{Pr(S_0 = s, A_0 = a, S_1 = s')}{Pr(A_0 = a, S_1 = s')} \\
&= \frac{d_0(s)\pi(s, a)P(s, a, s')}{\sum_{s_0} Pr(S_0 = s_0)Pr(S_1 = s', A_0 = a|S_0 = s_0)} \\
&= \frac{d_0(s)\pi(s, a)P(s, a, s')}{\sum_{s_0} d_0(s_0)\pi(s_0, a)P(s_0, a, s')}
\end{aligned}$$

.