Reinforcement Learning
Exercise 9 - Solution

Jonathan Schnitzler - st166934
Eric Choquet - st160996

July 7, 2024

REINFORCE on the Cart-Pole

a) Linear features Equation for $\pi(a|s,\theta)$:

$$\begin{aligned}
\pi(a|s,\theta) &= \frac{e^{\theta_a^T s}}{\sum_{c=0}^{1} e^{\theta_c^T s}} && (1)\\
&= \frac{e^{\theta_a^T s}}{e^{\theta_a^T s} + e^{\theta_b^T s}} && (2)\\
&= \frac{1}{1 + e^{(\theta_b - \theta_a)^T s}} && (3)
\end{aligned}$$

where θ_a and θ_b are the indices of the two actions. The four continuous state variables are for this purpose simply enumerated, i.e. $s = \begin{pmatrix} s_1 & s_2 & s_3 & s_4 \end{pmatrix}^T$. Differentiating with respect to a single parameter θ_i, we get

$$\begin{aligned}
\frac{\partial \pi(a|s,\theta)}{\partial \theta_i} &= \frac{s_i e^{(\theta_b - \theta_a)^T s}}{(1 + e^{(\theta_b - \theta_a)^T s})^2} && (4)\\
&= s_i \frac{1}{1 + e^{(\theta_b - \theta_a)^T s}} \left(1 - 1 + \frac{e^{(\theta_b - \theta_a)^T s}}{1 + e^{(\theta_b - \theta_a)^T s}} \right) && (5)\\
&= s_i \pi(a|s,\theta)(1 - \pi(a|s,\theta)) && (6)
\end{aligned}$$

Combining this for all parameters, we get the gradient of the policy

$$
\frac{\partial \pi(a|s,\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial \pi(a|s,\theta)}{\partial \theta_{a1}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{a2}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{a3}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{a4}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{b1}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{b2}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{b3}} \\ \frac{\partial \pi(a|s,\theta)}{\partial \theta_{b4}} \end{pmatrix} = \begin{pmatrix} s_1\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ s_2\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ s_3\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ s_4\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ -s_1\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ -s_2\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ -s_3\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ -s_4\pi(a|s,\theta)(1-\pi(a|s,\theta)) \end{pmatrix} = \begin{pmatrix} s\pi(a|s,\theta)(1-\pi(a|s,\theta)) \\ -s\pi(a|s,\theta)(1-\pi(a|s,\theta)) \end{pmatrix} \tag{7}
$$

b) Score function - Gradient of the log policy

$$
\log \pi(a|s,\theta) = \theta_a^T s - \log(e^{\theta_a^T s} + e^{\theta_b^T s}) \tag{8}
$$

Using the previous result, we can simply use the chain rule for the logarithmic function, for each variable respectively.

$$
\log(f(x))' = \frac{f'(x)}{f(x)} \tag{9}
$$

Thus, the gradient of the log policy cancels itself out

$$
\nabla \log \pi(a|s,\theta) = \begin{pmatrix} s(1-\pi(a|s,\theta)) \\ -s(1-\pi(a|s,\theta)) \end{pmatrix} \tag{10}
$$