

Exercise 08 [Avik Banerjee (3374885), Soumyadeep Bhattacharjee (3375428)]

(1) REINFORCE on the Cart-Pole

$$(a) \quad \pi(a|s, \theta) = \frac{e^{\theta^T s}}{\sum_b e^{\theta^T s}}$$

In the given situation,

$$\pi(a|s, \theta) = \frac{e^{\theta_a^T s}}{e^{\theta_a^T s} + e^{\theta_b^T s}}, \text{ where } a \text{ and } b \text{ are the two actions}$$

Now, $s \in \mathbb{R}^4$

$$\theta = \begin{bmatrix} \theta_{a_1} & \theta_{a_2} & \theta_{a_3} & \theta_{a_4} \\ \theta_{b_1} & \theta_{b_2} & \theta_{b_3} & \theta_{b_4} \end{bmatrix}$$

$$\nabla_\theta \pi(a|s, \theta) = \frac{(e^{\theta_a^T s} + e^{\theta_b^T s}) \frac{\partial}{\partial \theta} (e^{\theta_a^T s}) - e^{\theta_a^T s} \frac{\partial}{\partial \theta} (e^{\theta_a^T s} + e^{\theta_b^T s})}{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}$$

$$= \frac{(e^{\theta_a^T s} + e^{\theta_b^T s}) e^{\theta_a^T s} \begin{bmatrix} s \\ 0 \end{bmatrix}^T - e^{\theta_a^T s} \begin{bmatrix} e^{\theta_a^T s} s^T \\ e^{\theta_b^T s} s^T \end{bmatrix}}{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}$$

$$= e^{\theta_a^T s} \begin{bmatrix} (e^{\theta_a^T s} + e^{\theta_b^T s}) s^T - e^{\theta_a^T s} s^T \\ 0 - e^{\theta_b^T s} s^T \end{bmatrix}$$

$$= \frac{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}$$

$$= \frac{e^{\theta_a^T s} \begin{bmatrix} e^{\theta_b^T s} & s^T \\ -e^{\theta_b^T s} & s^T \end{bmatrix}}{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}$$

$$= \frac{e^{\theta_a^T s} e^{\theta_b^T s} \begin{bmatrix} s^T \\ -s^T \end{bmatrix}}{(e^{\theta_a^T s} + e^{\theta_b^T s})^2}$$

$$= h(s, a, \theta) \cdot h(s, b, \theta) \begin{bmatrix} s^T \\ -s^T \end{bmatrix}$$

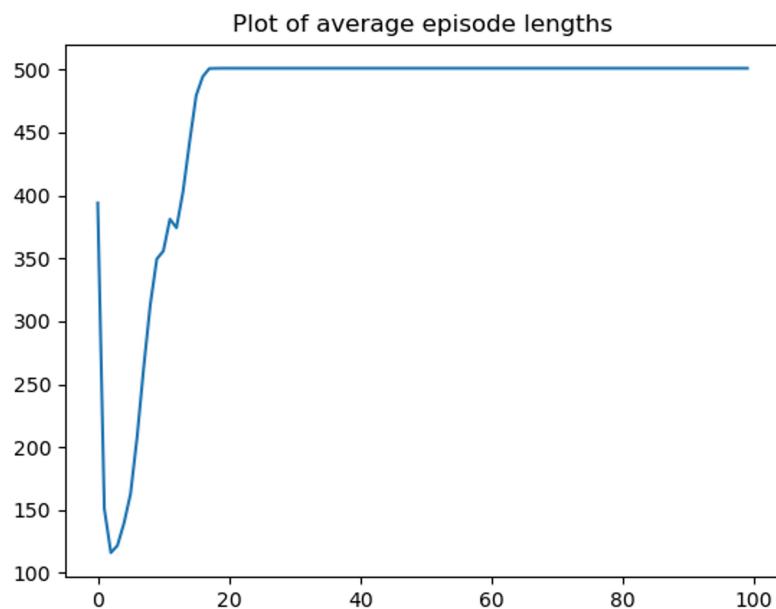
$$= h(s, a, \theta) \cdot h(s, b, \theta) \begin{bmatrix} s \\ -s \end{bmatrix}^T$$

$$\begin{aligned} (b) \log \pi(a|s, \theta) &= \log \frac{e^{\theta_a^T s}}{e^{\theta_a^T s} + e^{\theta_b^T s}} \\ &= \log(e^{\theta_a^T s}) - \log(e^{\theta_a^T s} + e^{\theta_b^T s}) \\ &= \theta_a^T s - \log(e^{\theta_a^T s} + e^{\theta_b^T s}) \end{aligned}$$

$$\begin{aligned} \nabla_{\theta} \log \pi(a|s, \theta) \\ = \nabla_{\theta} \left[\theta_a^T s - \log(e^{\theta_a^T s} + e^{\theta_b^T s}) \right] \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} s^T \\ 0 \end{bmatrix} - \frac{1}{(e^{\theta_a^T s} + e^{\theta_b^T s})} \begin{bmatrix} e^{\theta_a^T s} \cdot s^T \\ e^{\theta_b^T s} \cdot s^T \end{bmatrix} \\
&= \begin{bmatrix} s^T - \frac{e^{\theta_a^T s}}{e^{\theta_a^T s} + e^{\theta_b^T s}} s^T \\ 0 - \frac{e^{\theta_b^T s}}{e^{\theta_a^T s} + e^{\theta_b^T s}} s^T \end{bmatrix} \\
&= \begin{bmatrix} 1 - p(a|s, \theta) s^T \\ -p(b|s, \theta) s^T \end{bmatrix} \\
&= \begin{bmatrix} p(b|s, \theta) s^T \\ -p(b|s, \theta) s^T \end{bmatrix} \\
&= p(b|s, \theta) \begin{bmatrix} s^T \\ -s^T \end{bmatrix} = p(b|s, \theta) \begin{bmatrix} s \\ -s \end{bmatrix}^T
\end{aligned}$$

(c) Plot of average episode lengths :



(d) Ways to improve performance :

- A baseline can be subtracted from the obtained return while calculating the gradient, to reduce variance
- The returns of each step can be normalized by subtracting the mean and dividing by the standard deviation of the returns at all steps within the episode
- A neural network can be used to estimate the action probabilities at each state