CME241 Assignment16

Quinn Hollister

February 2022

1 Policy Gradient Actor-Critic Math

2 Evaluate the score function:

$$\begin{aligned}
\nabla_\theta \log \pi(s,a;\theta) =& \nabla_\theta(\phi(s,a)^T \cdot \theta - \log \sum_{b\in\mathcal{A}} e^{\phi(s,b)^T\cdot\theta}) \\
&= \phi(s,a) - \frac{1}{\sum_{b\in\mathcal{A}} e^{\phi(s,b)^T\cdot\theta}} \cdot \sum_{b\in\mathcal{A}} \phi(s,b)e^{\phi(s,b)^T\cdot\theta} \\
&= \phi(s,a) - \frac{\sum_{b\in\mathcal{A}} \phi(s,b)e^{\phi(s,b)^T\cdot\theta}}{\sum_{b\in\mathcal{A}} e^{\phi(s,b)^T\cdot\theta}} \\
&= \phi(s,a) - \sum_{b\in\mathcal{A}} \phi(s,b)\cdot\pi(s,b;\theta) \\
&= \phi(s,a) - \mathrm{E}_{\pi(s,b;\theta)}[\phi(s,b)]
\end{aligned}$$

The reason why the fractional sum can be reduced is because of the following fact: for given a specific action c, we can state the equality:

$$\frac{e^{\phi(s,c)^T\cdot\theta}}{\sum_{b\in\mathcal{A}} e^{\phi(s,b)^T\cdot\theta}} = \pi(s,c;\theta) \tag{1}$$

3 Construct the Action-Value approximation so that the CFAT is satisfied:

We want to construct $Q(s,a;\theta)$ such that:

$$\nabla_{\mathbf{w}} Q(s,a;\mathbf{w}) = \nabla_\theta \log \pi(s,a,\theta) \tag{2}$$

which can be easily constructed by setting each individual feature function to be the ith derivative of the score of the policy:

$$\begin{aligned}
Q(s,a;\mathbf{w}) &= \sum_{i=1}^{n} \phi_i(s,a)\cdot w_i \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\theta_i} \log \pi(s,a;\theta)\cdot w_i
\end{aligned}$$

Thus it is easily seen that the following is true:

$$\frac{\partial Q}{\partial w_i} = \frac{\partial}{\partial\theta_i} \log \pi(s,a;\theta) \tag{3}$$

4 Show that $Q(s, a; \mathbf{w})$ has zero mean for any state:

$$
\begin{aligned}
\mathrm{E}_{\pi(s,a;\theta)}[Q(s,a;\mathbf{w})] &= \sum_{a\in\mathcal{A}} \pi(s,a;\theta)\cdot Q(s,a;\mathbf{w}) \\
&= \sum_{a\in\mathcal{A}} \pi(s,a,\mathbf{w})\cdot (\sum_{i=1}^{n} \frac{\partial}{\partial\theta_i}\log\pi(s,a;\mathbf{w})\cdot w_i) \\
&= \sum_{a\in\mathcal{A}} (\sum_{i=1}^{n} \frac{\partial}{\partial\theta_i}\pi(s,a,\theta)\cdot w_i) \\
&= \sum_{i=1}^{n} (\sum_{a\in\mathcal{A}} \frac{\partial}{\partial\theta_i}\pi(s,a;\theta))\cdot w_i \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\theta_i}(\sum_{a\in\mathcal{A}} \pi(s,a,\theta))\cdot w_i \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\theta_i}(1)\cdot w_i \\
&= \sum_{i=1}^{n} 0 * w_i \\
&= 0
\end{aligned}
$$

The switch in summation's is valid because we're just adding terms and regrouping, the following simplified example demonstrates the operation:

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_1}\pi(s,a_1;\theta)\cdot w_1 + \frac{\partial}{\partial\theta_2}\pi(s,a_1;\theta)\cdot w_1 + \frac{\partial}{\partial\theta_1}\pi(s,a_2;\theta)\cdot w_1 + \frac{\partial}{\partial\theta_2}\pi(s,a_2;\theta)\cdot w_1 \\
&= (\frac{\partial}{\partial\theta_1}\pi(s,a_1;\theta) + \frac{\partial}{\partial\theta_1}\pi(s,a_2;\theta))\cdot w_1 + (\frac{\partial}{\partial\theta_2}\pi(s,a_1;\theta) + \frac{\partial}{\partial\theta_2}\pi(s,a_2;\theta))\cdot w_2
\end{aligned}
$$