Our code for the REINFORCE and for the ACTOR-CRITIC-ELIGIBILITY-TRACES algorithms can be found in `RL-book/Assignment16/assignment16_code.py` file

We based ourselves from the pseudo-codes we had in the lecture slides.

CME 241 : Assignment 16: Problem 3 : Pablo Veyrat

Let us consider the softmax function on the linear combination of features to approximate the policy function.

$$\pi(s,a,\theta) = \frac{e^{\phi(s,a)^T\cdot\theta}}{\sum_{b\in A} e^{\phi(s,b)^T\cdot\theta}}$$

. Let us evaluate the score function $\nabla_\theta \log \pi(s,a;\theta)$

We have $\log \pi(s,a,\theta) = \phi(s,a)^T\cdot\theta - \log\left(\sum_{b\in A} e^{\phi(s,b)^T\cdot\theta}\right)$

This gives us : $\frac{\partial \log \pi(s,a;\theta)}{\partial\theta_k} = \phi_k(s,a) - \frac{\sum_{b\in A}\phi_k(s,b)\, e^{\phi(s,b)^T\cdot\theta}}{\sum_{b\in A} e^{\phi(s,b)^T\cdot\theta}} = \phi_k(s,a) - \sum_{b\in A}\underbrace{\left(\frac{e^{\phi(s,b)^T\cdot\theta}}{\sum_{b\in A} e^{\phi(s,b)^T\cdot\theta}}\right)}_{\pi(s,b;\theta)}\phi_k(s,b)$

Hence $= \phi_k(s,a) - \underbrace{\sum_{b\in A}\pi(s,b;\theta)\,\phi_k(s,b)}_{=\mathbb{E}_\pi(\phi_k(s,\cdot))}$

In the end, we get that $\underline{\nabla_\theta \log \pi(s,a;\theta) = \phi(s,a) - \mathbb{E}_\pi(\phi(s,\cdot))}$

. Let us now construct the action-value function approximation so that we satisfy the key constraint of the compatible function approximation theorem.

We saw in class that a simple way to do this is to set $Q(s,a;w)$ to be linear in its features

If we let the features be $\nabla_\theta \log \pi(s,a;\theta)$ then $Q(s,a;w) = w^T\nabla_\theta \log \pi(s,a;\theta)$, and we then get:

$$\underline{\nabla_w Q(s,a;w) = \nabla_\theta \log(\pi(s,a;\theta))}$$

. Let us now show that $Q(s,a;w)$ has zero mean for any state s:

For $s\in S$ $\mathbb{E}_\pi(Q(s,a;w)) = \sum_{a\in A}\pi(s,a;\theta)\,Q(s,a;w)$

$= \sum_{a\in A}\pi(s,a;\theta)\, w^T\nabla_\theta \log \pi(s,a;\theta)$; and we have $\nabla_{\theta_i}\log \pi(s,a,\theta) = \frac{1}{\pi(s,a;\theta)}\frac{\partial \pi(s,a;\theta)}{\partial\theta_i}$

Hence $\mathbb{E}_\pi(Q(s,a;w)) = \sum_{a\in A}\pi(s,a;\theta)\sum_{i=1}^m w_i \frac{1}{\pi(s,a;\theta)}\frac{\partial\pi(s,a;\theta)}{\partial\theta_i}$

By rearranging the sums:

$$\mathbb{E}_\pi(Q(s,a;w)) = \sum_{a\in A}\sum_{i=1}^m w_i \frac{\partial\pi(s,a;\theta)}{\partial\theta_i}$$

$= \sum_{i=1}^m w_i \frac{\partial}{\partial\theta_i}\left(\sum_{a\in A}\pi(s,a;\theta)\right)$ and $\sum_{a\in A}\pi(s,a;\theta) = 1$, so $\frac{\partial}{\partial\theta_i}(1) = 0$, and this gives us:

$$\mathbb{E}_\pi(Q(s,a;w)) = \sum_{i=1}^m w_i \cdot 0 = 0$$

Thus $\underline{\forall s\in N \quad \mathbb{E}_\pi(Q(s,a;w)) = 0}$