

Our code for the REINFORCE and for the ACTOR-CRITIC-ELIGIBILITY-TRACES algorithms can be found in `RL-book/Assignment16/assignment16_code2.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.

$$\pi(s, a, \theta) = \frac{e^{\phi(s, a)^T \theta}}{\sum_{b \in A} e^{\phi(s, b)^T \theta}} \quad \text{let us evaluate the score function } \nabla_{\theta} \log \pi(s, a; \theta)$$

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

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

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

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

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 let $Q(s, a; w)$ 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:

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

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

$$\begin{aligned} \mathbb{E}_{\pi} (Q(s, a; w)) &= \sum_{a \in A} \pi(s, a; \theta) Q(s, a; w) \\ &= \sum_{a \in A} \pi(s, a; \theta) w^T \nabla_{\theta} \log \pi(s, a; \theta) \quad \text{and we have } \nabla_{\theta_i} \log \pi(s, a; \theta) = \frac{1}{\pi(s, a; \theta)} \frac{\partial \pi(s, a; \theta)}{\partial \theta_i} \end{aligned}$$

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

By rearranging the sums:

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

$$= \sum_{i=1}^m w_i \frac{\partial}{\partial \theta_i} \left(\sum_{a \in A} \pi(s, a; \theta) \right)$$

$$\text{and } \sum_{a \in A} \pi(s, a; \theta) = 1, \text{ so } \frac{\partial}{\partial \theta_i} (1) = 0, \text{ and this}$$

gives us:

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

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