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

$\pi(s,b;\theta)$

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 $\nabla_\theta \log \pi(s,a;\theta) = \phi(s,a) - \mathbb{E}_\pi(\phi(s,\cdot))$

- 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:

$\nabla_w q(s,a;w) = \nabla_\theta \log(\pi(s,a;\theta))$

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

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

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

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

You will find the code for this question in the `RL-book/Assignment2/assignment2_code.py` file.

We created a class to define the state. This class which we called `StateSnakeAndLadder` only has one attribute: it is the position.

The structure of the transition probabilities can be found in the function `get_transition_map` of the `StateSnakeAndLadderGame` class.

The graph of the probability distribution of time steps to finish the game is can be found here:

We used the code in the `RL-book/Assignment2/assignment2_code.py` file to generate it.

Equation example

$$\mathbb{P}(W \leq \theta_\gamma) = \gamma \infty \sum_{i=1}^{n} B_i \times exp(\mu + \sigma Z_i) \bar{\tau} \frac{\bar{Y}_{2^{n_0}}}{\bar{\tau}_{2_0^n}} \epsilon$$

```
test code
```

Example item

- **Step 1**:
- **Step 2**:
- **Step 3**:
- **Step 4**:

Code photo: