Assignment 16

Miao-Chin Yen

March 8, 2022

Problem 3

Assume we have a finite action space $\mathcal{A}$. Let $\boldsymbol{\phi}(s,a) = (\phi_1(s,a), \phi_2(s,a), \dots, \phi_m(s,a))$ be the features vector for any $s \in \mathcal{N}, a \in \mathcal{A}$. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$ be an m-vector of parameters. Let the action probabilities conditional on a given state s and given parameter vector $\boldsymbol{\theta}$ be defined by the softmax function on the linear combination of features: $\boldsymbol{\phi}(s,a)^T \cdot \boldsymbol{\theta}$, i.e.,

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

- Evaluate the score function $\nabla_{\boldsymbol{\theta}} \log \pi(s,a;\boldsymbol{\theta})$

- Construct the Action-Value function approximation $Q(s,a;\boldsymbol{w})$ so that the following key constraint of the Compatible Function Approximation Theorem (for Policy Gradient) is satisfied:

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

  where $\boldsymbol{w}$ defines the parameters of the function approximation of the Action-Value function.

- Show that $Q(s,a;\boldsymbol{w})$ has zero mean for any state s, i.e. show that

  $$\mathbb{E}_\pi[Q(s,a;\boldsymbol{w})] \text{ defined as } \sum_{a \in \mathcal{A}} \pi(s,a;\boldsymbol{\theta}) \cdot Q(s,a;\boldsymbol{w}) = 0 \text{ for all } s \in \mathcal{N}$$

Answer:

$$\log \pi(s,a;\boldsymbol{\theta}) = \boldsymbol{\theta} \cdot \boldsymbol{\phi}(s,a)^T - \log(\sum_{b \in \mathcal{A}} e^{\boldsymbol{\phi}(s,b)^T \cdot \theta})$$

$$\begin{aligned}
\frac{\partial \log \pi(s,a;\boldsymbol{\theta})}{\partial \theta_i} &= \phi_i(s,a) - \frac{\sum_{b \in \mathcal{A}} \phi_i(s,b) \cdot e^{\boldsymbol{\phi}(s,b)^T \cdot \theta}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\phi}(s,b)^T \cdot \theta}} \\
&= \phi_i(s,a) - \sum_{b \in \mathcal{A}} \frac{e^{\boldsymbol{\phi}(s,b)^T \cdot \theta}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\phi}(s,b)^T \cdot \theta}} \cdot \phi_i(s,b) \\
&= \phi_i(s,a) - \sum_{b \in \mathcal{A}} \pi(s,b;\boldsymbol{\theta}) \cdot \phi_i(s,b) \\
&= \phi_i(s,a) - \mathbb{E}_\pi\left[\phi_i(s,\cdot)\right]
\end{aligned}$$

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

Construct the Action-Value function approximation as follows:

$$Q(s,a;\boldsymbol{w}) = \boldsymbol{w}^T \cdot \nabla_\theta \log \pi(s,a,\boldsymbol{\theta})$$

Then we can satisfy the key constraint of the Compatible Function Approximation Theorem

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

And,

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot Q(s, a; \boldsymbol{w}) &= \sum_{a \in \mathcal{A}} \pi(s, a; \boldsymbol{\theta}) \cdot \boldsymbol{w}^T \cdot \nabla_\theta \log \pi(s, a, \boldsymbol{\theta}) \\
&= \sum_{a \in \mathcal{A}} \boldsymbol{w}^T \cdot \nabla_\theta \pi(s, a, \boldsymbol{\theta}) \\
&= \boldsymbol{w}^T \cdot \nabla_\theta \left(\sum_{a \in \mathcal{A}} \pi(s, a, \boldsymbol{\theta}) \right) \\
&= \boldsymbol{w}^T \cdot \nabla_\theta 1 \\
&= \boldsymbol{w}^T \cdot \mathbf{0} \\
&= 0
\end{aligned}$$