Carolyn Kao (chkao831@stanford.edu)

CME 241, Winter 2021

HW16

3. Let the action probabilities conditional on a given state s and given parameter vector θ be defined by the softmax function on the linear combination of features: $\boldsymbol{\phi}(s,a)^T \cdot \boldsymbol{\theta}$, i.e., $\pi(s,a;\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\phi}(s,a)^T \cdot \boldsymbol{\theta}}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\phi}(s,b)^T \cdot \boldsymbol{\theta}}}$

Evaluate the score function

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

$\Rightarrow \frac{\partial \log \pi(a|s;\theta)}{\partial \theta_i} = \phi_i(s,a) - \frac{\sum_{b \in \mathcal{A}} \phi_i(s,b) \cdot e^{\theta^T \cdot \phi(s,b)}}{\sum_{b \in \mathcal{A}} e^{\theta^T \cdot \phi(s,b)}}$

$= \phi_i(s,a) - \sum_{b \in \mathcal{A}} \pi(b \mid s;\theta) \cdot \phi_i(s,b)$

$= \phi_i(s,a) - \mathbb{E}_\pi \left[\phi_i(s,\cdot) \right]$ which further yields

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

Construct the Action-Value function approximation

Denote features of $Q(s,a;w)$ be $\nabla_\theta \log \pi(a \mid s,\theta)$

Let $Q(s,a;w)$ be linear in $Q(s,a;w) = w^T \cdot \nabla_\theta \log \pi(a \mid s,\theta)$ where w defines the parameters of the function approximation of the Action-Value function.

Show that Q has zero mean for any state s

$\sum_{a \in \mathcal{A}} \pi(a \mid ; \theta) \cdot Q(s,a;w)$

$= \sum_{a \in \mathcal{A}} w^T \cdot \nabla_\theta \pi(a \mid s,\theta) = w^T \cdot \nabla_\theta \left(\sum_{a \in \mathcal{A}} \pi(a \mid s,\theta) \right)$

$= w^T \cdot \nabla_\theta 1$

$= 0$