Worksheet 8

Stoyan Dimitrov
July 7, 2020

Policy Gradient

Exercise 1

a With $\pi(a|s,\theta) = \frac{e^{\theta_a^T s}}{\sum_k e^{\theta_k^T s}}$:

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}\pi(a|s,\theta) &= \sum_k \frac{\partial \pi(a|s,\theta)}{\partial \theta_k^T s}\frac{\partial \theta_k^T s}{\partial \theta_i} \\
&= \sum_k \frac{\partial \pi(a|s,\theta)}{\partial \theta_k^T s}\delta_{ik} s \\
&= \sum_k \pi(a|s,\theta)(\delta_{ak} - \pi(k|s,\theta))\delta_{ik} s \\
&= \pi(a|s,\theta)(\delta_{ai} - \pi(i|s,\theta))s
\end{aligned} \tag{1}$$

b

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}\log\pi(A_t|S_t,\theta) &= \frac{\frac{\partial}{\partial \theta_i}\pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} \\
&= \frac{\pi(A_t|S_t,\theta)(\delta_{ai} - \pi(i|S_t,\theta))S_t}{\pi(A_t|S_t,\theta)} \\
&= (\delta_{ai} - \pi(i|S_t,\theta))S_t
\end{aligned} \tag{2}$$

where $\delta_{ai} = 1$ if $a = i$, else 0
Then:

$$\nabla_\theta \log\pi(A_t|S_t,\theta) = \left[\frac{\partial}{\partial \theta_1}\log\pi(A_t|S_t,\theta), ..., \frac{\partial}{\partial \theta_n}\log\pi(A_t|S_t,\theta)\right] \tag{3}$$

c I'm pretty confident my gradient from b) is right. But applying each of the partial derivatives from the gradient on the corresponding dimension of θ doesn't seem to work well - I see some slow improvement. I'm also pretty confident that the update rule and the equation for $\nabla_\theta \log\pi(A_t|S_t,\theta)$ are different than in the Sutton book - we have matrix for θ_i and the same feature representation for every action and they have the same vector θ and different feature representation of the observation for each action. So the gradient w.r.t θ for them will be a vector and for us it's a matrix. My code can be found under https://github.com/StoyanVenDimitrov/rl-course/tree/master/ex08-pg. I'll be glad about some feedback what I'm doing wrong.

d By making use of the value function too: REINFORCE with baseline, actor-critic, ...