Prof. Joschka Boedecker, Gabriel Kalweit, Maria Huegle, Andreas Saelinger

Reinforcement Learning
Solution 10

1 Score Function

Assume a task with two discrete actions 0 and 1. Instead of a Gaussian policy or a softmax, we can define the policy to follow a Bernoulli distribution by the sigmoid function $\sigma(\cdot)$ over a linear combination of state features s and parameters θ, i.e. $\pi(a=1|s) = \sigma(s^T\theta)$ and $\pi(a=0|s) = 1-\sigma(s^T\theta)$.

Derive the score function for this policy.
Hint: the derivative of the sigmoid function is $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$.

Solution. The score function is defined as $\nabla_\theta \log \pi_\theta(s)$.

So for the defined policy:

$$
\begin{aligned}
\nabla_\theta \log \pi_\theta(a=1|s) &= \nabla_\theta \log \sigma(s^T\theta) \\
&= \frac{1}{\sigma(s^T\theta)}\sigma(s^T\theta)(1-\sigma(s^T\theta))s \\
&= (1-\sigma(s^T\theta))s \\
&= (1-\pi_\theta(a=1|s))s
\end{aligned} \tag{1}
$$

and

$$
\begin{aligned}
\nabla_\theta \log \pi_\theta(a=0|s) &= \nabla_\theta \log(1-\sigma(s^T\theta)) \\
&= \frac{1}{(1\sigma(s^T\theta))}(1-\sigma(s^T\theta))\sigma(s^T\theta)(-s) \\
&= -\sigma(s^T\theta)s \\
&= -\pi_\theta(a=1|s)s
\end{aligned} \tag{2}
$$