Policy Gradient Theorem

Behzad Khamidehi

To assess the behaviour of the policy gradient algorithms, we need to consider a performance measure. Since we study the episodic case here, the value of the start state is considered as the performance measure, i.e.,

$$J(\theta) = v_{\pi_\theta}(s_0) \tag{1}$$

Policy Gradient Theorem: The policy gradient theorem indicates that for any differentiable policy function $\pi_\theta(a|s)$, the gradient of function $J(\theta)$ can be expressed as

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi\{Q_{\pi_\theta}(s,a)\nabla_\theta \ln \pi_\theta(a|s)\} \tag{2}$$

Proof: To prove, we start from the state value function $v_\pi(s)$, $\forall s \in \mathcal{S}$, using the definition of the state value function, we can write

$$v_{\pi_\theta}(s) = \nabla_\theta \left[\sum_a \pi_\theta(a|s)Q_{\pi_\theta}(s,a)\right] \qquad \forall s \in \mathcal{S}. \tag{3}$$

Using the product rule, 3 can be written as

$$v_{\pi_\theta}(s) = \sum_a \left[\nabla_\theta \pi_\theta(a|s)Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\nabla_\theta Q_{\pi_\theta}(s,a)\right]. \tag{4}$$

Due to the fact that $Q_{\pi_\theta}(s,a) = \sum_{s',r} p(s',r|s,a)(r + v_{\pi_\theta}(s'))$, we have

$$v_{\pi_\theta}(s) = \sum_a \left[\nabla_\theta \pi_\theta(a|s)Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\nabla_\theta \sum_{s',r} p(s',r|s,a)(r + v_{\pi_\theta}(s'))\right]. \tag{5}$$

As mentioned before, the gradient is with respect to θ. However, $p(s',r|s,a)$ does not depend on θ. As a result, $\nabla_\theta \sum_{s',r} p(s',r|s,a)r = 0$, and

$$v_{\pi_\theta}(s) = \sum_a \left[\nabla_\theta \pi_\theta(a|s)Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\sum_{s'} p(s'|s,a)\nabla_\theta v_{\pi_\theta}(s')\right]. \tag{6}$$

By unrolling 6, we have

$$\begin{aligned} v_{\pi_\theta}(s) = \sum_a \Bigg[\nabla_\theta \pi_\theta(a|s)Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\sum_{s'} p(s'|s,a)\sum_{a'} \Big[\nabla_\theta \pi_\theta(a'|s')Q_{\pi_\theta}(s',a') + \\ \pi_\theta(a'|s')\sum_{s''} p(s''|s',a')\nabla_\theta v_{\pi_\theta}(s'')\Big]\Bigg]. \end{aligned} \tag{7}$$

If we repeat unrolling for k times and denote $\Pr(s \longrightarrow x, k, \pi)$ as the probability of transitioning from state s to state x in k steps under policy π, 7 can be rewritten as

$$v_{\pi_\theta}(s) = \sum_{x\in\mathcal{S}}\sum_{k=0} \Pr(s \longrightarrow x, k, \pi)\sum_a \nabla_\theta \pi_\theta(a|x)Q_{\pi_\theta}(x,a) \tag{8}$$

Using the resulting equation, we can write

$$\nabla_\theta J(\theta) = v_{\pi_\theta}(s_0) = \sum_{s \in \mathcal{S}} \left(\sum_{k=0} \Pr(s_0 \longrightarrow s, k, \pi) \right) \sum_a \nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s,a) \tag{9}$$

If we denote $\eta(s) = \sum_{k=0} \Pr(s_0 \longrightarrow s, k, \pi)$, we have

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \eta(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s,a) = \left(\sum_s \eta(s) \right) \sum_s \frac{\eta(s)}{\sum_s \eta(s)} \sum_a \nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s,a)$$

or equivalently,

$$\nabla_\theta J(\theta) \propto \sum_s \mu(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s,a), \tag{10}$$

where $\mu(s) = \frac{\eta(s)}{\sum\limits_s \eta(s)}$ is the stationary state distribution.

Since $\nabla_x \ln f(x) = \frac{\nabla_x f(x)}{f(x)}$, the last equation can be written as

$$= \sum_s \mu(s) \sum_a \pi_\theta(a|s) Q_{\pi_\theta}(s,a) \nabla_\theta \ln \pi_\theta(a|s) \tag{11}$$

$$\begin{aligned}
\nabla_\theta J(\theta) &\propto \sum_s \mu(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q_{\pi_\theta}(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \\
&= \sum_s \mu(s) \sum_a \pi_\theta(a|s) Q_{\pi_\theta}(s,a) \nabla_\theta \ln \pi_\theta(a|s) \\
&= \mathbb{E}_{s \sim \mu(s), a \sim \pi_\theta(a|s)} \left\{ Q_{\pi_\theta}(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right\} \\
&= \mathbb{E}_\pi \left\{ Q_{\pi_\theta}(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right\},
\end{aligned} \tag{12}$$

where $s \sim \mu(s), a \sim \pi_\theta(a|s)$ implies that both state and action distributions follow policy π_θ.

References

[1] Sutton, R. & Barto, A. Reinforcement Learning: An Introduction (MIT Press, 1998)