Paper Summary

Yu Chen

The Chinese University of HongKong, Department of Mechanical and Automation Engineering

E-mail: anschen@link.cuhk.edu.hk

Contents

1 Policy Gradient Methods for Reinforcement Learning with Function Approximation

This article[1] show theoretical possibility of using approximator to encode a policy. In this paper, the author prove the results for both two value functions, one is average reward,

$$V^{\pi}(s) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[r_1 + r_2 + \ldots + r_n|s,\pi] \tag{1.1}$$

$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{t=1}^{+\infty} r_t - V^{\pi}(s)|s,a,\pi], \tag{1.2}$$

the other is state-by-state reward,

$$V^{\pi}(s) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t|s,\pi], \tag{1.3}$$

$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}|s_t = s, a_t = a,\pi] \tag{1.4}$$

We use parameters θ to approximate policy π, which means $\pi(s,a) = \pi(s,a;\theta)$.

Theorem 1.0.1 (Policy Gradient). *For any MDP, in either the average-reward or state-state formulatios,*

$$\frac{\partial V^{\pi}(s)}{\partial \theta} = \int dsda\rho^{\pi}(s)\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a), \tag{1.5}$$

where $\rho^{\pi}(s)$ is discounted density for state s, defined as,

$$\rho^{\pi}(s) = \lim_{t\to\infty} p(s_t|s_0,\pi), \text{stationary distribution for average reward}, \tag{1.6}$$

$$\rho^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t p(s_t = s|s_0,\pi), \mathrm{state-state\ formulation}. \tag{1.7}$$

Proof:

Lemma 1.0.1. *The state value function can be written as,*

$$V^{\pi}(s) = \int dsda\rho^{\pi}(s)\pi(s,a)R_s^a, \tag{1.8}$$

where $R_s^a = \int ds' rp(r,s'|s,a)$, where $p(r,s'|s,a)$ is transition probability of MDPs.

Now, we firstly prove average-reward scheme.

$$
\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &= \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) && (1.9)\\
&= \sum_a \partial_\theta \pi Q^\pi + \pi \partial_\theta Q^\pi && (1.10)\\
&= \sum_a \partial_\theta \pi Q^\pi + \pi \partial_\theta \left(R_s^a - V^\pi(s) + \sum_{s'} p(s'|s,a) V^\pi(s') \right) && (1.11)\\
&= \sum_a \left(\partial_\theta \pi Q^\pi + \pi \sum_{s'} p(s'|s,a) \partial_\theta V^\pi(s') \right) - \partial_\theta V^\pi(s). && (1.12)
\end{aligned}
$$

Since, for averaged reward, the value function depends on the final stationary distribution, we can,

$$
\begin{aligned}
\sum_s \rho^\pi(s) \partial_\theta V^\pi(s) =& \sum_{s,a} \rho^\pi(s) Q^\pi(s,a) \partial_\theta \pi(s,a) + \sum_{s,a,s'} p(s'|s,a) \pi(s,a) \rho^\pi(s) \partial_\theta V^\pi(s') \\
&- \sum_s \rho^\pi(s) \partial_\theta V^\pi(s) \\
=& \sum_{s,a} \rho^\pi(s) Q^\pi(s,a) \partial_\theta \pi(s,a) + \sum_{s'} \rho^\pi(s') \partial_\theta V^\pi(s') - \sum_s \rho^\pi(s) \partial_\theta V^\pi(s) \\
=& \sum_{s,a} \rho^\pi(s) Q^\pi(s,a) \partial_\theta \pi(s,a).
\end{aligned} \tag{1.13}
$$

QED.

For state-state formulation, we can compute as,

$$
\begin{aligned}
\partial_\theta V^\pi(s) &= \sum_a \partial_\theta \pi Q^\pi + \pi \partial_\theta Q^\pi && (1.14)\\
&= \sum_a \partial_\theta \pi Q^\pi + \pi \partial_\theta \sum_{s'} r p(r, s'|s,a) + \gamma p(s'|s,a) V^\pi(s') && (1.15)\\
&= \sum_a \partial_\theta \pi(s,a) Q^\pi(s,a) + \pi(s,a) \gamma p(s'|s,a) \partial_\theta V^\pi(s') && (1.16)\\
&\ddots && (1.17)\\
&= \sum_s \rho^\pi(s) \sum_a Q^\pi(s,a) \partial_\theta \pi(s,a). && (1.18)
\end{aligned}
$$

From the above, we can find gradient of policy depends only on $Q^\pi(s,a)$ and is independent of $\partial_\theta Q^\pi$, which brings greate convinience. Then, we can introduce another approximation with parameters ω to approximate Q value function, denoted by $f_w(s,a)$. To make $f_w \to Q$, the following equation should be satisfied,

$$
\begin{aligned}
&\partial_w \sum_s \sum_a \rho^\pi(s,a) \pi(s,a) (Q(s,a) - f_w(s,a))^2 = 0 \\
\Leftrightarrow &\sum_s \sum_a \rho^\pi(s,a) \pi(s,a) (Q(s,a) - f_w(s,a)) \partial_w f_w(s,a) = 0.
\end{aligned} \tag{1.19}
$$

Theorem 1.0.2 (Policy Gradient with Function Approximation)**.** *If f_ω satisfies Eq.(1.19) and,*

$$\frac{\partial f_w(s,a)}{\partial \omega} = \frac{\partial \pi(s,a)}{\partial \theta}\frac{1}{\pi(s,a)}, \tag{1.20}$$

then,

$$\frac{\partial V^\pi(s)}{\partial \theta} = \sum_s \rho^s(\pi) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a). \tag{1.21}$$

The proof is obivous. In the following, we can discuss Eq.(1.20), and construct approximators satisfying it. For example, we use a Gibbs distribution to parameterize $\pi(s,a)$ as,

$$\pi(s,a;\theta) = \frac{e^{\theta^T \phi(s,a)}}{\sum_{a'} e^{\theta^T \phi(s,b)}}. \tag{1.22}$$

And then, we obtain,

$$\frac{1}{\pi(s,a)}\frac{\partial \pi(s,a)}{\partial \theta} = \phi(s,a) - \sum_b \pi(s,b)\phi(s,b) \tag{1.23}$$

Hence, we just need to parameterize $f_w(s,a)$ as,

$$f_w(s,a) = w^T(\phi(s,a) - \sum_b \pi(s,b)\phi(s,b)). \tag{1.24}$$

Theorem 1.0.3 (Policy Iteration with Function Approxmimation)**.** *Let π and f_w be any differential function approximators for the policy and value function respectively that satisfy the compatibility condition1.20 and for which $\max_{\theta,s,a,i,j} |\frac{\partial^2 \pi(s,a)}{\partial \theta_I \partial \theta_j}| < B < \infty$. Let $\{\alpha_k\}_{k=0}^{+\infty}$ be any step-size sequence such that $\lim_{k\to\infty} \alpha_k = 0$ and $\sum_k \alpha_k = \infty$. Then for any MDP with bounded rewards, the sequence $\{V^{\pi_k}\}$, defined by any θ_0, π_k, and,*

$$w_k = w \text{ such that } \sum_s \rho^{\pi_k} \sum_a \pi_k(s,a)(Q^{\pi_k}(s,a) - f_w(s,a))\partial_w f_w(s,a) = 0, \tag{1.25}$$

$$\theta_{k+1} = \theta + \alpha_k \sum_s \rho^{\pi_k}(s) \sum_a f_{w_k}(s,a)\partial_\theta \pi_k(s,a), \tag{1.26}$$

converges such that $\lim_{k\to\infty} \frac{\partial V^{\pi_k}(s)}{\partial \theta} = 0$

References

[1] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.