Policy Gradient Methods for Reinforcement Learning with Function Approximation

Zion Sung
Department of Computer Science
National Yang Ming Chiao Tung University
zionsung.c@nycu.edu.tw

1 Introduction

Function approximation is essential to reinforcement learning (RL). Such as neural networks which is used of generalizing function approximators. In previous work, the dominant approach is value-function approach. The value-function approach has worked well in many applications, but has several limitations. First, it is oriented toward finding deterministic policies, whereas the optimal policy is often stochastic. Second, an arbitrarily small change in the estimated value of an action can cause it to choose, or not to choose. Because of the discontinuous changes, it is hard to converge following the value-function approach. Below will discribe their three main contributions in the paper, and also compare to the prior works.

Firstly, they explore an new approach to function approximation in RL. Rather than approximating a value function using to compute a deterministic policy, they approximate a stochastic policy directly using an independent function approximator with its own parameters. Let θ denote the vector of policy parameters and ρ the performance of the policy (e.g., the average reward per step). Then, in the policy gradient approach, the policy parameters are updated approximately proportional to the gradient:

$$\Delta\theta \approx \alpha \frac{\partial \rho}{\partial \theta} \tag{1}$$

where α is a step size. If the above can be achieved, then θ can be asuured to converge to a locally optimal policy in terms of the performance measure ρ. Unlike the value-function approach, here small changes in θ can cause only small changes in the policy.

Secondly, they prove that an unbias estimate of the gradient (1) can be obtained from experience using an approximate value function satisfying certain properties. Willianms [1988], Williams [1992]REINFORCE algorithm also finds an unbiased estimate of the gradient, but without the assistance of a learned value function. Learning a value function and using it to reduce the variance of the gradient estimate appears to be important for rapid learning. Jaakkola et al. [1994] proved a results quite similar to this paper for the spacial case of function approximation corresponding to tabular POMDPs. This paper strengthens theirs and generalizes it to arbitrary differentiable function approximators.

Finally, they suggest a way of proving the convergence of a wide variety of algorithm based on policy-iteration architectures. Furthermore, they take the first step in this direction by proving that policy iteration with general differentiable function approximation is convergent to a locally optimal policy. In previous work, Baird and Moore [1998] obtained a weaker but similar result, their methods includes separately parameterized policy and value functions updated by gradient mehtods just like pocliy gradient. However, the methods do not converge to a locally optimal policy.

In my perspective, their methods proved to be converge to a locally optimal policy is outbreaking but also an important result in the realm of RL. While previous works in the period of time for a decade focus on estimating value using to find the deterministic policies. Until this paper proposed the method which can approximate a stochastic policy directly using an independent function approximator with

its own parameters, in addition, the small changes in policy parameres cause only small changes in the policy to make the policy more robust.

2 Problem Formulation

The authors consider a Markov decision process (MDP). The state, action, and reward at each time $t \in \{0, 1, 2, ...\}$ are denoted $s_t \in S, a_t \in A$, and $r_t \in R$ respectively. The environment's dynamics are defined by state transition probabilities, $P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$, and expected rewards $R^a_s = E\{r_{t+1} | s_t = s, a_t = a\}, \forall s, \in S, a \in A$,where $\theta \in R^l$ for $l \ll |S|$, is a parameter vector. Assume that π is differentiable with respect to its parameter, i.e., that $\frac{\partial \pi(s,a)}{\partial \theta}$ exists, also write just $\pi(s, a)$ for $\pi(s, a, \theta)$.

With function approximation, the authors use two ways of formulating the agent's objective. One is the average reward formulation, in which policies are ranked according to their long-term expected reward per step, $\rho(\pi)$

$$\rho(\pi) = \lim_{n\to\infty} \frac{1}{n} E\{r_1 + r_2 + \cdots + r_n | \pi\} = \sum_s d^\pi(s) \sum_a \pi(s, a) R^a_s,$$

where $d^\pi(s) = \lim_{t\to\infty} Pr\{s_t = s | s_0, \pi\}$ is the stationary distribution of states under π, which we assume exists and is independent of s_0 for all policies. In the average reward formulation, the value of a state-action pair given a policy is defined as

$$Q^\pi(s, a) = \sum_{n=1}^{\infty} E\{r_t - \rho(\pi) | s_0 = s, a_0 = a, \pi\}, \forall s \in S, a \in A.$$

The second formulation is that in which there is a designated start state s_0, and only care about the long-term reward obtained from it. The following are definitions of $\rho(\pi)$ and $Q^\pi(s, a)$.

$$\rho(\pi) = E\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \pi\}, Q^\pi(s, a) = E\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi\}$$

where $\gamma \in [0, 1]$ is a discount rate. In this formulation, the authors define $d^\pi(s)$ as a discounted weghting of states encountered starting at s_0 and then following π : $d^\pi(s) = \sum_{t=0}^{\infty} \gamma^t Pr\{s_t = s | s_0, \pi\}$.

3 Theoretical Analysis

In theoretical analysis, I will demonstrate three main theorem in this paper respectively. Includes proofs and some explanations.

3.1 Policy Gradient

Firstly, introduce policy gradient. For any MDP, in either the average-reward or start-state formulations,

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a). \tag{2}$$

The way of expressing the gradient was first discussed for the average-reward formulation by Marbach and Tsitsiklis [1999]. The authors of this paper extend their results to the start-state formulation and provide simpler and more direct proofs.

Proof:

$$\frac{\partial V^{\pi}(s)}{\partial \theta} := \frac{\partial}{\partial \theta}\sum_{a}\pi(s,a)Q^{\pi}(s,a), \forall s \in S$$
$$= \sum_{a}[\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \pi(s,a)\frac{\partial}{\partial \theta}Q^{\pi}(s,a)]$$
$$= \sum_{a}[\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \pi(s,a)[-\frac{\partial}{\partial \theta} + \sum_{s'}P^{a}_{ss'}\frac{V^{\pi}(s')}{\partial \theta}]]$$

Therefore,

$$\frac{\partial \rho}{\partial \theta} = \sum_{a}[\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \pi(s,a)\sum_{s'}P^{a}_{ss'}\frac{\partial V^{\pi}(s')}{\partial \theta}] - \frac{\partial V^{\pi}(s)}{\partial \theta}$$

Summing both sides over the stationary distribution d^{π},

$$\sum_{d}^{\pi}(s)\frac{\partial \rho}{\partial \theta} = \sum_{s}d^{\pi}(s)\sum_{a}\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \sum_{s}d^{\pi}(s)\sum_{a}\pi(s,a)\sum_{s'}P^{a}_{ss'}\frac{\partial V^{\pi}(s')}{\partial \theta}$$
$$- \sum_{s}d^{\pi}(s)\frac{\partial V^{\pi}(s)}{\partial \theta},$$

but since d^{π} is stationary,

$$\sum_{d}^{\pi}(s)\frac{\partial \rho}{\partial \theta} = \sum_{s}d^{\pi}(s)\sum_{a}\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \sum_{s}d^{\pi}(s)\frac{\partial V^{\pi}(s')}{\partial \theta} - \sum_{s}d^{\pi}(s)\frac{\partial V^{\pi}(s)}{\partial \theta},$$
$$\frac{\partial \rho}{\partial \theta} = \sum_{s}d^{\pi}(s)\sum_{a}\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a).$$

Q.E.D

For the start-state formulation:

$$\frac{\partial V^{\pi}(s)}{\partial \theta} := \frac{\partial}{\partial \theta}\sum_{a}\pi(s,a)Q^{\pi}(s,a), \forall s \in S$$
$$= \sum_{a}[\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \pi(s,a)\frac{\partial}{\partial \theta}[R^{a}_{s} + \sum_{s'}\gamma P^{a}_{ss'}V^{\pi}(s')]]$$
$$= \sum_{a}[\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a) + \pi(s,a)\sum_{s'}\gamma P^{a}_{ss'}V^{\pi}(s')]$$
$$= \sum_{x}\sum_{k=0}^{\infty}\gamma^{k}Pr(s \to x,k,\pi)\sum_{a}\frac{\partial \pi(x,a)}{\partial \theta}Q^{\pi}(x,a),$$

after serveral steps of unrolling, where $Pr(s \to x,k,\pi)$ is the probability of going from state s to state x in k steps under policy π. It is then immediate that

$$\frac{\partial \rho}{\partial \theta} = \frac{\partial}{\partial \theta}E\{\sum_{k=0}^{\infty}\gamma^{t-1}r_{t}|s_{0},\pi\} = \frac{\partial}{\partial \theta}V^{\pi}(s_{0})$$
$$= \sum_{s}\sum_{k=0}^{\infty}\gamma^{k}Pr(s_{0} \to s,k,\pi)\sum_{a}\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(x,a)$$
$$= \sum_{s}d^{\pi}(s)\sum_{a}\frac{\partial \pi(s,a)}{\partial \theta}Q^{\pi}(s,a).$$

Q.E.D

3.2 Policy Gradient with Approximation

Next, the center of this paper, they approximate Q^π by a learned function approximator. In prior works, Jaakkola et al. [1994] proved tha function approximation could ensure improment for moving in the direction in a tabular POMDP. Here, the authors extend their results to general function approximation and prove equality with the gradient. Let $f_w : SxA \to R$ be our approximation to Q^π, with parameter w. It is natural to learn f_w by following π and updating w by a rule such as $\triangle w_t \propto \frac{\partial}{\partial w}[\hat{Q}^\pi(s_t, a_t) - f_w(s_t, a_t)]^2 \propto [\hat{Q}^\pi(s_t, a_t) - f_w(s_t, a_t)]\frac{\partial f_w(s_t, a_t)}{\partial w}$, where $\hat{Q}^\pi(s_t, a_t)$ is some unbiased estimator of $Q^\pi(s_t, a_t)$, perhaps R_t. When such a process has converged to a local optimum, then

$$\sum_s d^\pi(s) \sum_a \pi(s, a)[Q^\pi(s, a) - f_w(s, a)]\frac{\partial f_w(s, a)}{\partial w} = 0 \tag{3}$$

Theorem 2 (Policy Gradient with Function Approximation).

If f_w satisfies (3) and is compatible with the policy parameterization in the same that

$$\frac{\partial f_w(s, a)}{\partial w} = \frac{\partial \pi(s, a)}{\partial \theta}\frac{1}{\pi(s, a)}, \tag{4}$$

then

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi s, a}{\partial \theta} f_w(s, a). \tag{5}$$

Proof: Combining (3) amd (4) gives

$$\sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta}[Q^\pi(s, a) - f_w(s, a)] = 0 \tag{6}$$

which tells us that the eroor in $f_w(s, a)$ is orthogonal to the gradient of the policy parameterization. Because the expression above is zero, we can subtract it from the policy gradient theorem (2) to yeild

$$\begin{aligned}
\frac{\partial \rho}{\partial \theta} &= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) - \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta}[Q^\pi(s, a) - f_w(s, a)] \\
&= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta}[Q^\pi(s, a) - Q^\pi(s, a) + f_w(s, a)] \\
&= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} f_w(s, a)
\end{aligned}$$

Q.E.D

3.3 Policy Iteration with Function Approximation

Finally, introduce policy iteration with function approximation. Given Theorem 2, we can prove for the first time that a form of policy iteration with function approxmation is convergent to a locally optimal policy. **Theorem 3 (Policy Iteration with Function Approximation)**. Let π and f_w be any differentiable function approximators for the policy and value function respectively that satisfy the compatibility condition (4) and for which $max_{\theta,s,a,i,j}|\frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j}| < B < \infty$. Let $\{a_k\}_{k=0}^\infty$ 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 $\{\rho(\pi_k)\}_{k=0}^\infty$, defined by any θ_0, $\pi_k = \pi(\cdot, \cdot, \theta_k)$, and

$$\begin{aligned}
w_k = w \text{ such that } &\sum_s d^{\pi_k}(s) \sum_a \pi_k(s, a)[Q^{\pi_k}(s, a) - f_w(s, a)]\frac{\partial f_w(s, a)}{\partial w} = 0 \\
= \theta_{k+1} = \theta_k + \alpha_k &\sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s, a)}{\partial \theta} f_{wk}(s, a),
\end{aligned}$$

converges such that $\lim_{k\to\infty} \frac{\partial \rho(\pi_k)}{\partial \theta} = 0$.

Proof: Theorem 2 assures that θ_k update is in the direction of the gradient. The bounds on $\frac{\partial^2_{\pi(s,a)}}{\partial \theta_i \partial \theta_j}$ and on the MDP's rewards together assure us that $\frac{\partial^2_{\rho}}{\partial \theta_i \partial \theta_j}$ is also bounded. These, together with the step-size requirements, are the necessary conditions to apply Proposition 3.5 from page 96 of Cao and Chen [1997], which assures convergence to a local optimum. Q.E.D

4 Conclusion

The theorem this paper proposed, in my perspective, is a turning point of the RL. It opens the new door of RL from approximating a value-function to approximating a stochastic policy, the polciy gradient. Furthermore, an extension of this method is also proposed, the policy gradient with approximation and policy iteration with function approximation. However, the these approach do not guaranteed monotonic improvement. In these paper, it does not discuss how to set the time-step α, which will be a significant impact on the robust of policy improment.

On 2015 Schulman et al. [2015] proposed the Trust Region Policy Optimization which guaranteed monotonic improvement. On 2017 the new methods Schulman et al. [2017], called proximal policy optimization (PPO), have some of the benefits of trust region policy optimization (TRPO), but they are much simpler to implement, more general, and have better sample complexity. These policy-gradient algorithms are on-policy by design; however, as on-policy algorithms, they suffer from poor sample efficiency. On 2018 Abdolmaleki et al. [2018] Maximum a posteriori policy optimisation (MPO) has proposed which are off-policy design. The mothod is highly data efficient, robust to hyperparameter choices and applicable to complex control problems.

As mentioned above, we can see how significant the impact of this theorem is. Of course there is still room for improvement. And it's worthwhile for us to explore.

References

RJ Willianms. Toward a theory of reinforcement-learning connectionist systems. *Technical Report NU-CCS-88-3, Northeastern University*, 1988.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Tommi Jaakkola, Satinder Singh, and Michael Jordan. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, 7, 1994.

Leemon Baird and Andrew Moore. Gradient descent for general reinforcement learning. *Advances in neural information processing systems*, 11, 1998.

Peter Marbach and John N Tsitsiklis. Simulation-based optimization of markov reward processes: implementation issues. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, volume 2, pages 1769–1774. IEEE, 1999.

Xi-Ren Cao and Han-Fu Chen. Perturbation realization, potentials, and sensitivity analysis of markov processes. *IEEE Transactions on Automatic Control*, 42(10):1382–1393, 1997.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.