Policy Gradient Methods for Reinforcement Learning with Function Approximation

Pin-Hsuan Chiang
Institute of Data Science and Engineering
National Yang Ming Chiao Tung University
`angel0705.cs10@nycu.edu.tw`

1 Introduction

Approximate value functions to determine policies prove to be theoretically intractable. Value-based methods approximate the value function. The policy-based method is implicitly expressed, generally using a greedy algorithm, such as the policy of selecting the action with the highest estimated value in each state. Although the value-based method performs well in many applications, it still has some limitations. The first is that it is oriented to a deterministic policy, but the optimal policy is usually stochastic, and the selected action will correspond to a probability(Jaakkola et al. [1994]), which is difficult to achieve for value-based. Then, a very small change in the estimated value of an action may change the possibility of this action being selected. For example, the DQN series of methods Q-learning and Sarsa(Gordon [1996]) have insufficient processing ability for continuous actions. If you pick actions in a continuous space, then the Q-value will calculate the value in an infinite number of actions, and then choose the action. In the end, the algorithm can't converge.

This paper directly approximates stochastic policies using independent functions with its own parameters. For example, a policy might be a neural network whose inputs is state, whose output is action selection probabilities, and whose weights are policy parameters. ρ is the evaluation index of system performance. θ is parameters of the policy function. α is the positive definite step size. The gradient update is equal to α multiplied by θ and partially differentiated from ρ.

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

A method that directly approximates a stochastic policy, and the update value of the policy gradient is approximately proportional to the performance gradient. And if the above is implemented, then θ is generally guaranteed to converge to a local optimum policy in the performance measure ρ.

This paper proves that an unbiased estimate of the gradient can be obtained empirically by using an approximation function that satisfies some properties. Various algorithms based on "actor-critic" or policy iteration architectures are also proved to converge.

This paper focuses on deriving a method to directly approximate policy functions for continuous action spaces. First of all, it is determined that the update policy parameters of this method are to be updated according to the direction of the gradient of the objective function. Then different forms of the objective function are proposed, but the gradients of the different forms are consistent. The next step is to approximate the action value function Q. Finally, it is proved that the policy iteration of arbitrary differentiable function approximation can converge to the local optimum policy.

The advantages of the policy-based method are the first to be more efficient in continuous action spaces, and the second is a policy that can achieve stochastic. The last one is that in some cases, the value function may be more difficult to calculate, while the policy function is easier. The disadvantage is that it usually converges to a local optimum rather than a global optimum. And evaluating a strategy is often inefficient. Because the process can be slow, there are also many inefficient attempts and high variance.

2 Problem Formulation

Markov Decision Process (MDP) (Sutton and Barto [1998]):

$s_t \in \mathcal{S}$ is state at time t, where $\mathcal{S}$ is action space. $t \in \{0, 1, 2, ...\}$

s_0 is initial state.

$a_t \in \mathcal{A}$ is action at time t, where $\mathcal{A}$ is state space. $t \in \{0, 1, 2, ...\}$

$r_t \in \mathcal{R}$ is reward at time t, where $\mathcal{R}$ is reward function. $t \in \{0, 1, 2, ...\}$

$P^a_{ss'} = Pr\{s_{t+1} = s'|s_t = s, a_t = a\}$ is state transition probability, which is the probability of jumping to a state s' from the current state s and action a.

$R^a_s = E\{r_{t+1} = s'|s_t = s, a_t = a\}, \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$ is the expected reward over all the possible states that one can transition to from state s and action a.

$\pi(s, a, \theta) = Pr\{a_t = a|s_t = s, \theta\}, \forall s \in \mathcal{S}, a \in \mathcal{A}$, is policy defines the thought behind picking an action given the current state s. where $\theta \in R^l$ for $l \ll |\mathcal{S}|$is a parameter vector. For convenience, write $\pi(s, a)$ for $\pi(s, a, \theta)$. And assume $\frac{\partial \pi(s,a)}{\partial \theta}$ exists.

$\gamma \in [0, 1]$ is discount factor, using a discount is that there is no certainty about the future rewards.

State-value function:

$$V^\pi(s) = \sum_a \pi(s, a) Q^\pi(s, a)$$

3 Theoretical Analysis

3.1 Policy Gradient Theorem

In MDP, state transition depends not only on the state but also on the policy. First define ρ as the average reward of the current policy, independent of the initial state, the initial state is random, and in the no discount mode is

$$\rho(\pi) = \lim_{n\to\infty} \frac{1}{n} E\{r_1 + r_2 + ... + r_n|\pi\}$$

The stationary distribution of the Markov process, when the transition probability satisfies certain conditions, the distribution of the final state will tend to be stable. Assuming that the MDP converges before a finite constant N, then it will enter a stationary distribution, and the corresponding average cumulative sum is:

$$d^\pi(s) = \lim_{n\to\infty} \Pr\{s_t = s|s_0, \pi\}$$

Therefore, the stationary state distribution of states under π, $d^\pi(s)$ generated by the policy is used to weight $\rho(\pi)$, as follows

$$\rho(\pi) = \sum_s d^\pi(s) \sum_a \pi(s, a) \mathcal{R}^a_s$$

so,

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

Then based on $\rho(\pi)$, start from state s and take action a, and then follow policy π, the Action-value function Q for policy π is defined as:

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

For the MDP with a fixed initial state s_0, then define ρ and Q in another way, they have discount and give only one result:

$$\rho(\pi) = E\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t|s_0, \pi\}$$

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

Given action a and follow policy π, the weighted sum of discounts for states that may be encountered is

$$d^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \Pr\{s_t = s | s_0, \pi\}$$

The discounted reward function is more widely used because it converges faster and is easier to compute.

In Policy Gradient, the goal is to make $\rho(\pi)$ as large as possible. Assuming π is differentiable with respect to parameter θ, then we have Theorem 1.

Theorem 1 (Policy Gradient). For any MDP, in either the average-reward or start -state for formulations,

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

Proof: Let's look at the one without discount first. State-value function for policy π is

$$V^\pi(s) = \sum_a \pi(s,a) Q^\pi(s,a)$$

And then partial differential to theta. The second to third equation is by Bellman formula $Q^\pi(s,a) = R_s^a - \rho(\pi) + \sum\limits_{s'} P_{ss'}^a V^\pi(s')$.

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &\overset{def}{=} \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) \quad \forall s \in \mathcal{S} \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} Q^\pi(s,a)] \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} [\mathcal{R}_s^a - \rho(\pi) + \sum_{s'} P_{ss'}^a V^\pi(s')]] \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) [-\frac{\partial \rho}{\partial \theta} + \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta}]]
\end{aligned}$$

Bring $\frac{\partial \rho}{\partial \theta}$ out because it has nothing to do with a, therefore,

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

Multiply by $\sum\limits_s d^\pi(s)$, summing both sides on a stationary distribution,

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

Because $d^\pi(s)$ is a stationary distribution, select an action, move to the next state, it is still a stationary distribution. Then $\sum\limits_s d^\pi(s) \sum\limits_a \pi(s,a) \sum\limits_{s'} P_{ss'}^a = \sum\limits_{s'} .d^\pi(s')$.Therefore,

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

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

And then there is the fixed initial state. The second to third equation is by Bellman formula $Q^\pi(s,a) = R_s^a + \sum\limits_{s'} \gamma P_{ss'}^a V^\pi(s')$.The four to last equation $Pr(s \to x, k, \pi)$ represents the probability of transitioning from state s to state x through k steps under policy π.

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &\overset{def}{=} \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a) \quad \forall s \in \mathcal{S} \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} Q^\pi(s,a)] \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} [\mathcal{R}_s^a + \sum_{s'} \gamma P_{ss'}^a V^\pi(s')]] \\
&= \sum_a [\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \sum_{s'} \gamma P_{ss'}^a V^\pi(s')] \\
&= \sum_x \sum_{k=0}^\infty \gamma^k Pr(s \to x, k, \pi) \sum_a \frac{\partial \pi(x,a)}{\partial \theta} Q^\pi(x,a)
\end{aligned}$$

If k = 0, then x = s, $Pr(s \to x, k, \pi) = 1$. That's $\sum\limits_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$.

If k = 1, then $Pr(s \to x, k, \pi) = \pi(s,a) \sum\limits_a P_{sx}^a$, and so on.

we can get $\rho(\pi) = V^\pi(s_0)$. Therefore,

$$\begin{aligned}
\frac{\partial \rho}{\partial \theta} &= \frac{\partial}{\partial \theta} E\{\sum_{t=1}^\infty \gamma^{t-1} r^t | s_0, \pi\} = \frac{\partial}{\partial \theta} V^\pi(s_0) \\
&= \sum_s \sum_{k=0}^\infty \gamma^k Pr(s_0 \to s, k, \pi) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) \\
&= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)
\end{aligned}$$

3.2 Policy Gradient with Approximation

Consider $Q^\pi(s,a)$ represented by a function approximation, in order to solve the real computational problem for policy gradients. Let function $f_w : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ be an approximation of $Q^\pi(s,a)$, its weight is w. The gradient of w is $\Delta w_t \propto \frac{\partial}{\partial w} [\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)]^2 \propto [\hat{Q}^\pi(s_t,a_t) - f_w(s_t,a_t)] \frac{\partial f_w(s_t,a_t)}{\partial w}$, where $\hat{Q}^\pi(s_t,a_t)$ is some biased estimator of $Q^\pi(s_t,a_t)$. Approximation of approximation function $f_w(s,a)$ to $Q^\pi(s,a)$ under policy π. If $f_w(s,a)$ converges to a local optimum, then the final convergence result should satisfy the following,

$$\sum_s d^\pi(s) \sum_a \pi(s,a) [Q^\pi(s,a) - f_w(s,a)] \frac{\partial f_w(s,a)}{\partial w} = 0 \tag{1}$$

Theorem 2 (Policy Gradient with Function Approximation). If $f_w(s,a)$ is already an exact approximation, then it satisfies (1). And it compatible with policy parameterization in a sense. If the following formula is satisfied,

$$\frac{\partial f_w(s,a)}{\partial w} = \frac{\partial log[\pi(s,a)]}{\partial \theta} = \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} \tag{2}$$

then,

$$\frac{\partial \rho}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a) \tag{3}$$

Proof: Combining (1) and (2), then

$$\sum_s d^\pi(s) \sum_a \pi(s,a)[Q^\pi(s,a) - f_w(s,a)] \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} = 0 \Rightarrow \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta}[Q^\pi(s,a) - f_w(s,a)] = 0 \tag{4}$$

We can subtract (4) from theorem 1 policy gradient: $\frac{\partial \rho}{\partial \theta} = \sum\limits_s d^\pi(s) \sum\limits_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$. Because (4) is equal to zero. After the calculation, we can get the proof.

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

3.3 Application to Deriving Algorithms and Advantages

We can use Theorem 2 to derive the approximate value function $f_w(s,a)$. There are many policies. We choose the policy function is linear softmax policies $\pi(s,a) = \frac{\exp^{\theta^T \phi_{sa}}}{\sum_b \exp^{\theta^T \phi_{sb}}}$, where each ϕ_{sa} is an l-dimensional feature vector about characterizing state-action pair (s, a).
So we substitute it into (2),

$$\begin{aligned}
\frac{\partial f_w(s,a)}{\partial w} &= \frac{\partial \pi(s,a)}{\partial \theta} \frac{1}{\pi(s,a)} \\
&= [\phi_{sa} - \sum_b \pi(s,b)\phi_{sb}] \frac{\exp^{\theta^T \phi_{sa}}}{\sum_b \exp^{\theta^T \phi_{sb}}} \frac{\sum_b \exp^{\theta^T \phi_{sb}}}{\exp^{\theta^T \phi_{sa}}} \\
&= \phi_{sa} - \sum_b \pi(s,b)\phi_{sb}
\end{aligned}$$

After integrating it, we can get

$$f_w(s,a) = w^t[\phi_{sa} - \sum_b \pi(s,b)\phi_{sb}]$$

This means that the approximation $f_w(s,a)$ to the state is a linear function of some features. It must be linear in the same features as the policy.

At the same time, it can be noticed that f_w satisfies the following formula,

$$\sum_a \pi(s,a) f_w(s,a) = 0, \forall s \in \mathcal{S}$$

f_w requires that each state has zero mean.

This means that $Q^\pi(s,a)$ may no longer be the best option and we should use the advantage function (Baird III [1993]).It is equal to action value function minus state value function, and it represents the advantage of an action a relative to the average in state s and action a.

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

We know that when $f_w(s,a)$ is a functional approximation of $Q^\pi(s,a)$, then

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

that is, Theorem 2 (3).

And then we can add a value function or an arbitrary state function that is an approximation to it.

$$\because \sum_a \pi(s,a) = 1 \quad \Rightarrow \sum_a \frac{\partial \pi(s,a)}{\partial \theta} = 0$$

$$\therefore \sum_a \frac{\partial \pi(s,a)}{\partial \theta} V^\pi(s) = 0$$

So we can get,

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

Therefore, $\frac{\partial \rho}{\partial \theta} = \sum\limits_s d^\pi(s) \sum\limits_a \frac{\partial \pi(s,a)}{\partial \theta} [f_w(s,a) + v(s)]$), where v: $\mathcal{S} \rightarrow \mathcal{R}$ is an arbitrary function.

3.4 Convergence of Policy Iteration with Function Approximation

Here we will prove that the iterative form of policy with function approximation converges to a local optimum policy by Theorem 2.

Theorem 3 (Policy Iteration with Function Approximation).

Let f_w and π be any differentiable function approximators of value and policy functions satisfying condition (4), for which $max_{\theta,s,a,i,j} |\frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j}| < B < \infty$, that it is finite. Here is the algorithm's requirements for the step size (learning rate) α. Let $\{a_k\}_{k=0}^\infty$ be any step-size sequence , so that we can get $\lim\limits_{k \rightarrow \infty} \alpha_k = 0$ and $\sum\limits_k \alpha_k = \infty$. Then, for any MDP with bounded rewards, any θ_0 define the sequence $\{\rho(\pi_k)\}_{k=0}^\infty$, policy function $\pi_k = \pi(\cdot, \cdot, \theta_k)$, $w_k = w$ so that

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

$$\theta_{k+1} = \theta_k + \alpha_k \sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$$

converges so that

$$\lim_{k \rightarrow \infty} \frac{\partial \rho(\pi_k)}{\partial \theta} = 0$$

Proof: By Theorem 2, we can get $\frac{\partial \rho(\pi_k)}{\partial \theta} = \sum\limits_s d^{\pi_k}(s) \sum\limits_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$, then gradient update is $\theta_{k+1} = \theta_k + \alpha_k \sum\limits_s d^{\pi_k}(s) \sum\limits_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a)$.

And then, $\frac{\partial^2 \pi(s,a)}{\partial \theta_i \partial \theta_j}$ has bound, and $\frac{\partial^2 \rho}{\partial \theta_i \partial \theta_j}$ also has bound.

Finally, the step size α has $\lim\limits_{k\to\infty} \alpha_k = 0$ and $\sum\limits_k \alpha_k = \infty$. Therefore, $\lim\limits_{k\to\infty} \frac{\partial\rho(\pi_k)}{\partial\theta} = 0$, it converges to a local optimum, by Proposition 3.5(Convergence for a Diminishing Stepsize) from page 96 of Bertsekas and Tsitsiklis.

4 Conclusion

This paper writes three theorems about policy gradients. Theorem 1 is the policy gradient. And if the approximation process of the action-value function Q by the approximation function f_w in a certain learning process converges to the local optimum, then there is formula (1), and the approximate value function f_w and the policy function pi satisfy formula (4), then there is formula (5), which is also Theorem 2. The final theorem 3 proves that only when there are restrictions on the step size, the policy gradient of any differentiable function approximation can converge to a local optimum policy.

Policy-Based applies to a continuous action space.And it will not converge to a deterministic value, it will tend to generate optimal stochastic policy. In Value-Base, a very small change in the approximation of an action of the value function may change the probability of this action being selected, but Policy-Based avoids this problem.

References

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

Geoffrey J Gordon. Chattering in sarsa (lambda)-a cmu learning lab internal report. 1996.

Richard S Sutton and Andrew G Barto. Reinforcement learning: an introduction mit press. *Cambridge, MA*, 22447, 1998.

Leemon C Baird III. Advantage updating. Technical report, WRIGHT LAB WRIGHT-PATTERSON AFB OH, 1993.

Dimitri P Bertsekas and JN Tsitsiklis. Neuro-dynamic programming, athena scientific, belmont, ma, 1996. *Google Scholar Google Scholar Digital Library Digital Library*.