Policy Gradient Methods for Reinforcement Learning With Function Approximation and Actor Critic Algorithms

Anay Pattanaik, Aristomenis Tsopelakos, and Wyatt McAllister[1]

***Abstract*— Methods for value function approximation are a critical component of reinforcement learning algorithms. However, there are many theoretical challenges to standard approaches to value function approximation, which would allow an optimal policy to be computed. This work presents an alternate approach utilizing separate function approximation for both the policy and the value function. Here, the value function is updated along the gradient of the expected long term return for the current estimate of the policy. This is made possible by estimating the gradient from the state-action value function. The novel result here is that this policy iteration scheme is convergent to a locally optimal policy, given a specific class of differentiable function approximators. This theoretical result increases the applicability of actor-critic algorithms to high performance and safety critical domains, where convergence guarantees are needed.**

I. Introduction

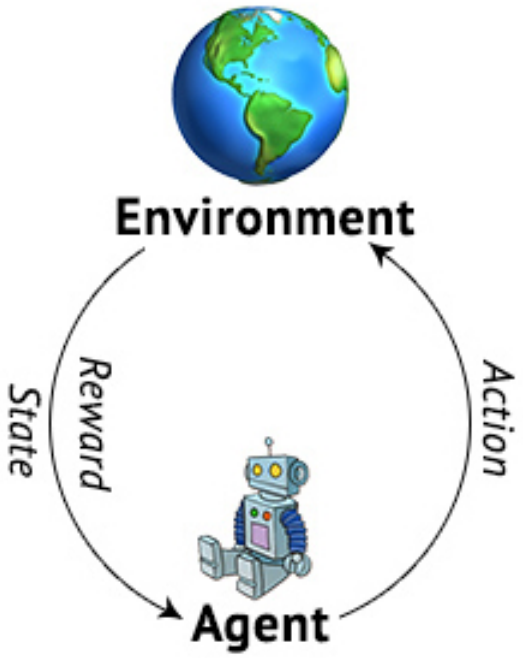

Fig. 1: Reinforcement Learning Framework

Reinforcement Learning is a process by which tasks are learned based on environmental stimulus. The learner chooses actions based on the perceived state of the environment to maximize a reward metric. This reward metric may be perceived or inferred directly from the environmental state. The agent observes the consequences of its actions on the reward metric through the environmental state, and optimizes its actions to yield the greatest long-term expected reward. Without knowing the underlying distribution of the data, the agent must plan an optimal policy using only the observations from interactions with the environment.

Past work on reinforcement learning has utilized a value-function based approach, in which the function approximation is used to determine the state-action value function for each state-action pair, and the policy is then selected greedily with respect to this value function. The limitations of this approach are that it selects a deterministic policy, which may not account for the stochastic nature of the real environment, and that small perturbations to the value function can have large effects on the policy.

This work examines an alternative method which directly approximates a stochastic policy via an independent function approximation scheme. This approximation scheme is based on policy parameters, θ, and on the average reward per step, ρ, where $\alpha > 0$ is a fixed step size.

$$\theta_{t+1} \leftarrow \theta_t + \alpha \nabla_\theta \rho \tag{1}$$

The above scheme converges to a locally optimal policy under the measure of ρ. Furthermore, for this approach, small changes in θ will on cause small changes in the policy.

It is shown that the policy gradient may be computed utilizing a separate estimate for the value function satisfying certain key characteristics. Algorithms of this type are known as actor-critic algorithms. Here, the "actor" is the policy planned by the agent, which governs its actions, and the "critic" is the state-action value function, which critiques the policy. Other schemes for estimation of the policy gradient without computing the value function exist, but they converge much more slowly. The value function estimate expedites computation by reducing the variance in the estimate for the policy gradient. This work presents convergence results which are valid for all function approximators within a specific class, where previously, convergence results were not available. This theoretical result increases the usefulness of actor-critic algorithms in high performance and safety critical domains, which require convergence guarantees.

*This work was done towards completing the project component of ECE543

[1]Anay Pattanaik is with the Dept. of Computer Science, University of Illinois at Urbana-Champaign.
Aristomenis Tsopelakos and Wyatt McAllister are with the Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign walt@illinois.edu, wmcalli22@illinois.edu

II. Policy Gradient Theorem

Consider the following reinforcement learning problem, which is presented as a tuple of state s_t, action a_t, and reward r_t, at each time t.

$$s_t \in S, \quad a_t \in A, \quad r_t \in R$$

The state transition probability, which governs the evolution of the system from one state to another given a proposed action, is as follows.

$$P^a_{ss'} = P\left(s_{t+1} = s' \middle| s_t = s, a_t = a\right)$$

The expected reward is given below. As stated previously, this is the quantity the learner attempts to maximize over time through its choice of action.

$$R^a_s = E\left[r_{t+1} | s_t = s, a_t = a\right], \forall s, s' \in S, a \in A$$

At each step of the decision process, the agent plans a policy, $\pi(s, a, \theta)$, which maps the current state s, to an action a, based on a set of parameters given by θ, where $\pi(s, a, \theta)$, $\theta \in R^l$ for $l \ll |S|$. This policy is chosen to ensure that the agents action at any state will maximize the long-term expected return.

$$\pi(s, a, \theta) = \pi_\theta(s, a) = P\left(a_t = a | s_t = s, \theta\right), \forall s \in S, a \in A$$

We assume the policy is differentiable.

$$\pi(s, a, \theta) \in C^1 : \nabla_\theta \pi(s, a, \theta) \text{ exists}$$

We assume that for any starting state, s_0, there exists a unique stationary distribution, $d^{\pi_\theta}(s)$, independent of s_0. In other words, eventually the probability of visiting any state under a given policy is constant in time.

$$d^{\pi_\theta}(s) = \lim_{t\to\infty} P\left(s_t = s | s_0, \pi_\theta\right)$$

We define the long-term expected reward per step as $\rho(\pi_\theta)$. Maximizing this quantity is equivalent to solving the reinforcement learning problem.

$$\rho(\pi_\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E\left[r_t | \pi_\theta\right] = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) R^a_s$$

We define the state-action value function, $Q^{\pi_\theta}(s, a)$, which assigns a valuation to any state-action pair, allowing us to evaluate any candidate policy. Conceptually, we want to change the policy in a manner which increases $Q^{\pi_\theta}(s, a)$.

$$Q^{\pi_\theta}(s, a) = \sum_{t=1}^{\infty} E\left[r_t - \rho(\pi_\theta) | s_0 = s, a_0 = a, \pi_\theta\right]$$

Our goal is use our state-action value function, $Q^{\pi_\theta}(s, a)$, to update the parameters, θ, which uniquely determine the policy, $\pi(s, a, \theta)$, in a manner which increases the long-term expected reward, $\rho(\pi_\theta)$. Thus, we want to show mathematically that the rate of change of $\rho(\pi_\theta)$ with respect to θ is positively correlated with the rate of change of $\pi(s, a, \theta)$ with respect to θ, and with $Q^{\pi_\theta}(s, a)$. The Policy Gradient Theorem, stated next, does exactly that.

Theorem II.1. *For any MDP,*

$$\nabla_\theta \rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(s, a) Q^{\pi_\theta}(s, a) \tag{2}$$

Proof. In, [1], the proofs for both the above average reward formulation, as well as an analogous start-state formulation, are considered. For simplicity, we consider only the average reward formulation here, as they are both equivalent. First, define the state value function, $V^{\pi_\theta}(s)$, as below.

$$V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s, a)$$

Then, the following equations hold:

$$\rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) R^a_s$$

$$Q^{\pi_\theta}(s, a) + \rho(\pi_\theta) = R^a_s + \sum_{s'} P^a_{ss'} V^{\pi_\theta}(s')$$

where $d^\pi(s) = \lim_{t\to+\infty} P[s_t = s | s_o; \pi]$ We first differentiate the above expression for $V^{\pi_\theta}(s)$.

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(s) &= \nabla_\theta \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \\
&= \sum_a \left[\nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a) + \pi_\theta(a|s) \nabla_\theta Q^{\pi_\theta}(s, a)\right]
\end{aligned}$$

We then substitute the expression for $Q^{\pi_\theta}(s, a)$.

$$\begin{aligned}
&= \sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \\
&+ \sum_a \pi_\theta(a|s) \left[\sum_{s'} P^a_{ss'} \nabla_\theta V^{\pi_\theta}\left(s'\right) - \nabla_\theta\left(\rho(\pi_\theta)\right)\right]
\end{aligned}$$

Since $\nabla_\theta\left(\rho(\pi_\theta)\right)$ is independent of a, we can take it outside the sum, which becomes one, as $\pi_\theta(a|s)$ is a probability distribution.

$$\begin{aligned}
&= \sum_a \left[\nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a)\right] \\
&+ \sum_a \left[\pi_\theta(a|s) \sum_{s'} P^a_{ss'} \nabla_\theta V^{\pi_\theta}\left(s'\right)\right] - \nabla_\theta \rho_\theta(\pi_\theta)
\end{aligned}$$

We multiply both sides of the previous equation by stationary distance $d^{\pi_\theta}(s)$, and sum it over all s.

$$\begin{aligned}
&\sum_s d^{\pi_\theta}(s) \nabla_\theta V^{\pi_\theta}(s) \\
&= \sum_s d^{\pi_\theta}(s) \sum_a \left[\nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a)\right] \\
&+ \sum_s d^{\pi_\theta}(s) \sum_a \left[\pi_\theta(a|s) \sum_{s'} P^a_{ss'} \nabla_\theta V^{\pi_\theta}\left(s'\right)\right] \\
&- \sum_s d^{\pi_\theta}(s) \nabla_\theta \rho(\pi_\theta)
\end{aligned}$$

We observe that $d^{\pi_\theta}(s)$ is independent of a and s', so we bring it inside.

$$= \sum_s d^{\pi_\theta}(s) \sum_a \left[\nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)\right]$$
$$+ \sum_s \sum_a \left[\pi_\theta(a|s) \sum_{s'} d^{\pi_\theta}(s)\, P^a_{ss'} \nabla_\theta V^{\pi_\theta}(s')\right]$$
$$- \sum_s d^{\pi_\theta}(s)\, \nabla_\theta \rho(\pi_\theta)$$

Since, $d^{\pi_\theta}(s)$ is a stationary distribution, $d^{\pi_\theta}(s)\, P^a_{ss'}$ is equal to $d^{\pi_\theta}(s')$.

$$= \sum_s d^{\pi_\theta}(s) \sum_a \left[\nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)\right]$$
$$+ \sum_s \sum_a \left[\pi_\theta(a|s) \sum_{s'} d^{\pi_\theta}(s') \nabla_\theta V^{\pi_\theta}(s')\right]$$
$$- \sum_s d^{\pi_\theta}(s)\, \nabla_\theta \rho(\pi_\theta)$$

Now, since $d^{\pi_\theta}(s')\, \nabla_\theta V^{\pi_\theta}(s')$ is independent of a and s, we interchange the order of the sums.

$$= \sum_s d^{\pi_\theta}(s) \sum_a \left[\nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)\right]$$
$$+ \sum_{s'} d^{\pi_\theta}(s') \nabla_\theta V^{\pi_\theta}(s') \sum_s \sum_a \pi_\theta(a|s)$$
$$- \sum_s d^{\pi_\theta}(s)\, \nabla_\theta \rho(\pi_\theta)$$

Finally, since $\pi_\theta(a|s)$ is a probability distribution, the two sums go to one.

$$= \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)$$
$$+ \sum_{s'} d^{\pi_\theta}(s') \nabla_\theta V^{\pi_\theta}(s') - \nabla_\theta \rho(\pi_\theta)$$

Therefore, after canceling $\sum_{s'} d^{\pi_\theta}(s') \nabla_\theta V^{\pi_\theta}(s')$ on both sides, we get the desired result.

$$\nabla_\theta \rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a)$$

■

III. Policy Gradient with function Approximation

In the previous section, we assumed that we had knowledge of the state-action value function, $Q^{\pi_\theta}(s,a)$, which was utilized to update the parameters, θ. However, in reality, this information may not actually be available to us. Thus, we would like to estimate $Q^{\pi_\theta}(s,a)$, and show that our previous update scheme for the parameters, θ, will still yield an optimum long-term expected reward $\rho(\pi_\theta)$, even when the estimate for $Q^{\pi_\theta}(s,a)$ is utilized in place of its actual value. One may think that the performance of the algorithm would be affected by the choice of function approximator. Here, we show that the previous algorithm is still optimal when the estimate of $Q^{\pi_\theta}(s,a)$ is used, under any function approximation scheme within the class of interest.

Now, we will consider the case in which $Q^{\pi_\theta}(s,a)$ is given by another differential function approximator, which we denote $f_w : S \times A \to R$, with w being another vector of parameters with dimension m, estimated via the following gradient update. Here, $\hat{Q}^{\pi_\theta}(s_t, a_t)$ is an unbiased estimator for $Q^{\pi_\theta}(s_t, a_t)$.

$$\Delta w_t \sim \frac{\partial}{\partial w}\left[Q^{\pi_\theta}(s_t,a_t) - f_w(s_t,a_t)\right]^2$$
$$\sim \left[Q^{\pi_\theta}(s_t,a_t) - f_w(s_t,a_t)\right] \nabla_w f_w(s_t,a_t)$$

The convergence criteria for this process is as follows.

$$\sum_{s,a} d^{\pi_\theta}(s)\, \pi_\theta(a|s)\, \left[Q^{\pi_\theta}(s,a) - f_w(s,a)\right] \nabla_w f_w(s,a) = 0$$

Theorem III.1. *If f_w satisfies:*

$$\sum_{s,a} d^{\pi_\theta}(s)\pi_\theta(a|s)[Q^{\pi_\theta}(s,a) - f_w(s,a)]\nabla_w f_w(s,a) = 0 \quad (3)$$

and is compatible with the policy parametrization in the sense that:

$$\nabla_w f_w(s,a) = \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \quad (4)$$

Then,

$$\nabla_\theta \rho(\pi_\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a|s) f_w(s,a) \quad (5)$$

Proof. In, [1], (3) is subtracted from the Policy Gradient Theorem to yield the result. In our proof, we instead utilize the potential function, E, inspired by the potential functions used in proofs of stochastic gradient descent.

$$E = \frac{1}{2} \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, \left[f_w(s,a) - Q^{\pi_\theta}(s,a)\right]^2$$

We differentiate E with respect to w.

$$\nabla_w E = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, f_w(s,a)\, \nabla_w f_w(s,a)$$
$$- \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a)\, \nabla_w f_w(s,a) = 0$$

Using the above expression for $\nabla_w f_w(s,a)$, we have the following.

$$\sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, f_w(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}$$
$$- \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} = 0$$

We rearrange terms.

$$\sum_s d^{\pi_\theta}(s) \sum_a f_w(s,a) \nabla_\theta \pi_\theta(a|s)$$
$$= \sum_s d^{\pi_\theta}(s) \sum_a Q^{\pi_\theta}(s,a)\, \nabla_\theta \pi_\theta(a|s)$$

From Theorem II.I, we have the following.

$$\sum_s d^{\pi_\theta}(s) \sum_a Q^{\pi_\theta}(s,a)\, \nabla_\theta \pi_\theta(a|s) \overset{Theorem II.I}{=} \nabla_\theta \rho(\pi_\theta)$$

■

IV. Application to Deriving Algorithms and Advantages

We will use the previous theorem to derive an appropriate form of the parameterization for the value function. We consider the following policy parameterization, composed of a linear combination of features, called the Gibbs distribution.

$$\pi_\theta(s,a) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi b}}, \forall s \in S, a \in A$$

Here, each ϕ_{sa} is a feature vector of dimension l for (s,a). By condition (4), we have the following.

$$\nabla_w f_w(s,a) = \nabla_\theta \pi_\theta(s,a) \frac{1}{\pi_\theta(s,a)} = \phi_{sa} - \sum_b \pi_\theta(s,b) \phi_{sb}$$

Therefore, a natural parametrization of f_w is the following.

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

This parameterization is a linear combination of the same features as in the policy parameterization, but is normalized to have mean zero at every state. These relations describe the class of function approximators used in this work.

$$\sum_a \pi_\theta(s,a) f_w(s,a) = 0, \forall s \in S$$

V. Convergence of Policy Iteration with Function Approximation

Theorem V.1. *Let π, and f_w, be any differentiable function approximators for the policy, and value function, respectively that satisfy the compatibility condition:*

$$\nabla_w f_w(s,a) = \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}$$

and for which $\max_{\theta,s,a,i,j} \left\| \frac{\partial^2 \pi_\theta(s,a)}{\partial \theta_i \partial \theta_j} \right\| < B < +\infty$. *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* $\{\rho(\pi_k)\}_{k=0}^{\infty}$, *defined by any* θ_0, $\pi_k = \pi(.,.,\theta_k)$, *and* $w_k = w$ *such that*

$$\sum_{s,a} d^{\pi_k}(s) \pi_k(s,a) \left[Q^{\pi_k}(s,a) - f_w(s,a) \right] \nabla_w f_w(s,a) = 0$$

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

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

Proof. By Theorem III.1, θ_k updates in the direction of the gradient. Furthermore, we know that both $\frac{\partial^2 \pi_\theta(s,a)}{\partial \theta_i \partial \theta_j}$, and the reward for the MDP, are bounded. These facts, together with the above restrictions on the step size, tell us that $\frac{\partial \rho(\pi_k)}{\partial \theta}$ converges to a local optimum by Proposition 3.5 given on Page 96 of [3]. ∎

VI. Actor-Critic Algorithms

Generally, the class of reinforcement learning algorithms which use a separate estimate for the state-action value function to update the estimate for the policy are called actor-critic algorithms. Here, the actor is the policy estimate, which affects the system, and the critic is the state-action value function estimate, which evaluates the policy. In these algorithms, the weight update rule for the parameterization of the critic is updated via the gradient as follows.

$$w_{t+1} \leftarrow w_t + \Delta w_t$$

The previous sections considers a gradient law for w_t which is written explicitly below. The temporal difference (TD), is $\left(Q_{t+1}^{\pi_\theta}(s_t,a_t) - f_{w_t}^{\pi_\theta}(s_t,a_t) \right)$.

$$\Delta w_t = \alpha \left(Q_{t+1}^{\pi_\theta}(s_t,a_t) - f_{w_t}^{\pi_\theta}(s_t,a_t) \right) \nabla_w f_{w_t}^{\pi_\theta}(s_t,a_t)$$

An alternative version of this algorithm is called TD(λ), in which a weighted combination of the gradients $\nabla_w f_k^{\pi_\theta}(s_t,a_t)$ at different times k up to t is utilized.

$$\Delta w_t = \alpha \left(Q_{t+1}^{\pi_\theta}(s_t,a_t) - f_{w_t}^{\pi_\theta}(s_t,a_t) \right) \sum_{k=1}^{t} \lambda^{t-k} \nabla_w f_{w_k}^{\pi_\theta}(s_t,a_t)$$

Corollary VI.1. *For every $\varepsilon > 0$, there exists λ sufficiently close to 0, such that in the TD(λ) algorithm,*

$$\liminf_{k\to\infty} \nabla_\theta \rho(\pi_k) \overset{a.s}{\leq} \varepsilon$$

Proof. If λ is close to zero, then the algorithm reduces to that considered previously. Then, by Theorem V.1, $\frac{\partial \rho(\pi_k)}{\partial \theta}$ goes to zero in the limit as k goes to ∞. This is a stronger result than that which is claimed, so the corollary is proved.

$$\lim_{k\to\infty} \nabla_\theta \rho(\pi_k) = 0 \Rightarrow \liminf_{k\to\infty} \nabla_\theta \rho(\pi_k) \leq \varepsilon \quad \forall \varepsilon > 0$$

∎

VII. Conclusions

This work presents novel convergence results for an actor-critic method for reinforcement learning employing a TD(λ) gradient update scheme, where λ is close to zero. The results here hold any function approximator within a specific class, whereas previously, convergence results were not available. The faster convergence rate of actor-critic methods over previous methods for function approximation in reinforcement learning make them useful in a variety of domains. The novel convergence results presented here are of great benefit to the community, as they will allow these more efficient algorithms to be used in high-performance or safety-critical domains, which require convergence guarantees.

References

[1] Sutton, Richard S., et al. "Policy gradient methods for reinforcement learning with function approximation." Advances in neural information processing systems. 2000.

[2] Konda, Vijay R., and John N. Tsitsiklis. "Actor-critic algorithms." Advances in neural information processing systems. 2000.

[3] Bertsekas, D. P. "J.. N. Tsitsiklis." Neuro-dynamic Programming (1996).