Policy Gradient Methods for Reinforcement Learning with Function Approximation

explained by Uk Jo

03/20/19

- Richard S. Sutton wrote this paper!!!
- Old paper but it is a mandatory paper for understanding PG algorithm such as actor-critic PG variants including DDPG, A2C, A3C, TRPO, PPO and of course REINFORCE including every PG algorithms

1 Abstract

- Policy is explicitly represented by its own function approximator, independent of the value function(REINFORCE) and it is updated according to the gradient of expected reward with respect to the policy parameters(actor-critic).
- PG can be written in a form suitable for estimation from experience aided by an approximate action-value or advantage function.
- For the first time, it prove that a version of policy iteration with arbitrary differentiable function approximation is convergent to a locally optimal policy.

2 Limitation of DQN(value based)

The value-function approach has worked well in many applications, but has several limitations.

- deterministic policies << stochastic policies(Singh, Jaakkola, and Jordan, 1994)
- Discontinuous changes is a key obstacle for convergence(Bertsekas and Tsitsiklis, 1996).
- Q-Learning, Sarsa, and dynamic programming methods have all been shown unable to converge to any policy for simple MDPs and simple function approximators.

3 Policy gradient

A stochastic policy approximate to a function approximator(f) with its own parameters.

$$f(state) \approx Pr(action)$$

Let θ denote the vector of policy parameters ρ the performance of the corresponding policy

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

where α is a positive-definite step size. θ can usually be assured to converge to a locally optimal policy in the performance measure ρ. Unlike the value-function approach, here small changes in θ can cause only small changes in the policy and in the state-visitation distribution.

This paper will prove an unbiased estimate of the gradient can be obtained from experience using an approximate value function satisfying certain properties. (REINFORCE, without the assistance of a learned value function, but learns much more slowly than RL methods using value functions and has received relatively little attention)

4 Previous work

- Learning a value function and using it to reduce the variance of the gradient estimate appears to be essential for rapid learning.(Jaakkola, Singh and Jordan (1995))

- Our result also suggests a way of proving the convergence of a wide variety of algorithms based on "actor-critic" or policy-iteration architectures (e.g., Barto, Sutton, and Anderson, 1983; Sutton, 1984; Kimura and Kobayashi, 1998).

- Konda and Tsitsiklis (in prep.) independently developed a very simialr result to ours. See also Baxter and Bartlett (in prep.) and Marbach and Tsitsiklis (1998). Our result strengthens theirs and generalizes it to arbitrary differentiable function approximators.

- First version of policy iteration with general differentiable function approximation is convergent to a locally optimal policy.

5 Policy Gradient Theorem

Let's consider MDP and $\frac{\partial \pi(s,a)}{\partial \theta}$ exists.

- States $s_t \in \mathcal{S}$, Actions $a_t \in \mathcal{A}$, Rewards $r_t \in \mathcal{R}$,

- $\mathcal{P}^a_{ss'} = \mathcal{P}_r\{s_{t+1} = s' | s_t = s, a_t = a\}$,

- expected rewards $\mathcal{R}^a_s = E\{r_{t+1} | s_t = s, a_t = a\}, \forall s \in \mathcal{S}, a \in \mathcal{A}$

- $\pi(s, a, \theta) = \mathcal{P}r\{a_t = a | s_t = s, \theta\}, \forall s \in \mathcal{S}, a \in \mathcal{A}$, where $\theta \in \mathcal{R}^l, for$ $l <<$ |S|,is a parameter vector

5.1 Two objective function

- 1. Average reward formulation

  policies are ranked according to their long-term expected reward per step, $\rho(\pi)$:

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

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

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

- 2. Long-term reward formulation

  designated start state s_0, and we care only about the long-term reward obtained from it. We will give our results only once, but they will apply to this formulation as well under the definitions

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

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

Our first result concerns the gradient of the performance metric with respect to the policy parameter:

5.2 Theorem 1. Policy theorem

For any MDP, in either the average-reward or start-state formulations,

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

See the proof of theorem 1 in original paper.

- there are no terms of the form $\frac{\partial d^\pi(s)}{\partial \theta} Q^\pi(s,a))$: the effect of policy changes on the distribution of states does not appear, which means that it is convenient for approximating the gradient by sampling.
- For example, if s was sampled from the distribution obtained by following π, then $\sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)$ would be an unbiased estimate of $\frac{\partial \rho}{\partial \theta}$.

- $Q^{\pi}(s,a)$ is also not normally known and must be estimated. One approach is to use the actual returns, $R_t = \sum_{k=1}^{\infty} r_{t+k} - \rho(\pi)$ (or $R_t = \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}$ in the start-state formulation) as an approximation for each $Q^{\pi}(s_t, a_t)$. This leads to Williams's episodic REINFORCE algorithm,

6 Poilcy Gradient with Approximation

Now consider the case in which Q^{π} is approximated by a learned function approximator. If the approximation is sufficiently good, we might hope to use it in place of Q^{π} in (2) and still point roughly in the direction of the gradient. Previously, Jaakkola, Singh, and Jordan (1995) proved it in a tabular POMDP to assure positive inner product with the gradient, which is sufficient to ensure improvement for moving in that direction. Here we extend their result to general function approximation and prove equality with the gradient.

- Let $f_w : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ be our approximation to Q^{π} with parameter w.
- learn f_w by following π and updating w by a rule such as $\Delta w_t \propto \frac{\partial}{\partial w}[\hat{Q}^{\pi}(s_t, a_t) - f_w(s_t, a_t)]^2 \propto [\hat{Q}^{\pi}(s_t, a_t) - f_w(s_t, a_t)]\frac{\partial f_w(s,a)}{\partial w}$, 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

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

6.1 Theorem 2. Policy Gradient with Function Approximation

If f_w satisfies above and is compatible with the policy parameterization in the sense that

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

then

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

See proof of Theorem 2 in original paper.

7 Applications to Deriving Algorithms and Advantages

Given a policy parameterization, Theorem 2 can be used to derive an appropriate form for the value-function parameterization. For example, consider a policy that is a Gibbs distribution in a linear combination of features:

$$\pi(s,a) = e^{\theta^T \phi_{sa}} / \sum_b e^{\theta^T \phi_{sb}}$$

so that the natural parameterization of f_w is

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

In other words, f_w must be linear in the same features as the policy, except normalized to be mean zero for each state. Other algorithms can easily be derived for a variety of nonlinear policy parameterizations, such as multi layer backpropagation networks.

above for f_w requires that it have zero mean for each state: $\sum_a \pi(s,a) f_w(s,a) = 0, \forall s \in \mathcal{S}$. In this sense it is better to think of f w as an approximation of the advantage function, $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ (much as in Baird, 1993), rather than of Q^π. Our convergence requirement (3) is really that f_w get the relative value of the actions correct in each state, not the absolute value, nor the variation from state to state. Our results can be viewed as a justification for the special status of advantages as the target for value function approximation in RL.

In fact, our (2), (3), and (5), can all be generalized to include an arbitrary function of state added to the value function or its approximation. For example, (5) can be generalized to $\frac{\partial \rho}{\partial \theta} = \sum_a \frac{\partial \pi(s,a)}{\partial \theta [f_w(s,a) + v(s)]}$, where $v : \mathcal{S} \to \mathcal{R} isanarbitraryfunction.(Thisfollowsimmediatelybecause \sum_a \frac{\partial \pi(s,a)}{\partial \theta} = 0, \forall s \in \mathcal{S}$) The choice of v does not affect any of our theorems, but can substantially affect the variance of the gradient estimators. The issues here are entirely analogous to those in the use of reinforcement baselines in earlier work (e.g., Williams, 1992; Dayan, 1991; Sutton, 1984). In practice, v should presumably be set to the best available approximation of V7r. Our results establish that that approximation process can proceed without affecting the expected evolution of f_q and π.

8 Convergence of Policy Iteration with Function Approximation

Given Theorem 2, we can prove for the first time that a form of policy iteration with function approximation is convergent to a locally optimal policy.

8.1 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 comapatibility condition (4) and for which $max_{\theta,s,a,i,j}|\frac{\partial \pi(s,a)}{\partial \theta_i \partial \theta_j}| < B < \infty$.

- Let $a_k{}_{k=0}^{\infty}$ be any step-size sequence such that $lim_{k\to\infty} a_k = 0$. Then, for any MDP with bounded rewards, the sequence $\{\rho(\pi_k)\}_{k=0}^{\infty}$, defined by any $\theta_0, \pi_k = \pi(\cdot, \cdot, \theta_k)$, and

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

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

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

See this proof in orignal paper