A Note on Off-policy Actor Critic

Ian Lin
Department of Information Management and Finance
National Yang Ming Chiao Tung University
ianlienfa.dif07@nycu.edu.tw

1 Introduction

This paper is first published in 2012, and presented both theory and practical use of off-policy actor critic structure. At that time point, most of the researches have been done on on-policy settings. Off-policy TD learning had just gained some advances, with the proposed method, we can make use of the benefit of off-policy learning and also be able to utilize a larger action space.

We can apply policy gradient methods on the actor-critic framework, that is, directly learn the policy to avoid the downsides of action-value methods, including the limitations of non-stochastic policies, hard-to-solve maximization problems, and over-sensitive policies(small change on action-value can drastically changes the policy)

There are three major contributions of this paper:

- For the critic, the authors proposed an version of GTD(λ)
- The authors provided the details and the proof of the convergence
- Empirical experiences on ACER and other algorithms are given

The proposed method, Off-PAC algorithm, is shown to outperform many other off-policy methods such as $Q(\lambda)$, Greedy-GQ and Softmax-GQ for the "Continuous grid world" environment, and has similar or slightly better results with compared to other methods under other environments such as "Mountain Car" and "Pendulum".

2 Problem Formulation

We consider a Markov decision process with discrete state space S, discrete action space A and a distribution $P : S \times A \times S' \to [0,1]$ with transition probability $P(s'|s,a)$ and expected reward $R : S \times A \times S'$. For the process we see a sequence of data come in tuple form: $(s_t, a_t, {s'}_t, r_t)$ for $t = 1, 2, ...$, where $s_t \in S$, ${s'}_t \in S$, $a_t \in A$ and $r_t \in R$. Since this is an off-policy setting, so the action is determined by the behavior policy $b(a|s)$

For Actor-Critic methods, we also need the value and value-action function, we first denote termination condition as $d : S \to [0,1]$, which takes in a state and output 1 when termination state is reached. The value function for $\pi : S \times A \to (0,1]$ is:

$$V^{\pi,d} = E[r_{t+1}, r_{t+2}, ... + r_{t+T} | s_t = s], \forall s \in S \tag{1}$$

And the action-value function is:

$$Q^{\pi,\gamma}(s,a) = \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma(s')V^{\pi,\gamma}(s')] \tag{2}$$

for all $a \in A$ and $s \in S$

Our goal is to find a policy π_u: $A \times S \rightarrow [0,1]$ that with weight vector u and maximize:

$$J_\gamma(u) = \sum_{s' \in S} d^b(s) V^{\pi_u,\gamma}(s) \tag{3}$$

where $d^b(s) = \lim_{t\rightarrow\infty} P(s_t = s|s_0, b)$ (Notice that the objective is weighted by $d^b(s)$ since we are optimizing it under the behavior policy b)

3 Theoretical Analysis

For this paper, the critic is updated using the GTD(λ) algorithm, namely

$$MSPBE(v) = \|\hat{V} - \prod T_\pi^{\lambda,\gamma}\hat{V}\|_D^2 \tag{4}$$

For the policy gradient, they proposed the Off-policy Policy gradient theorem, here we take gradient on the last term of equation 3 and get:

$$\nabla_u J_\gamma(u) = \nabla_u[\sum_{s\in S} d^b(s)V^{\pi_u,\gamma}(s)] \tag{5.1}$$

$$\nabla_u J_\gamma(u) = \sum_{s\in S} d^b(s) \sum_{a\in A}[\nabla_u \pi(a|s) Q^{\pi,\gamma}(s,a) + \pi(a|s)\nabla_u Q^{\pi,\gamma}(s,a)] \tag{5.2}$$

The paper futher provided two justification, which are respectively "Policy Improvement" and "Off-policy Policy Gradient Theorem" to show that it is reasonable to omit the last term of the above equation:

$$\nabla_u J_\gamma(u) = \sum_{s\in S} d^b(s) \sum_{a\in A}[\nabla_u \pi(a|s) Q^{\pi,\gamma}(s,a)] \tag{5.3}$$

Next, to make use of samples from behavior policy to do the update, so equation 5.2 can be rewrite as:

$$\begin{aligned} g(u) &= E[\sum_{a\in A} \nabla_u \pi(a|s) Q^{\pi,\gamma}(s,a)|s \sim d^b] && (6.1)\\ &= E[\sum_{a\in A} b(a|s)\frac{\pi(a|s)}{b(a|s)}\frac{\nabla_u\pi(a|s)}{\pi(a|s)} Q^{\pi,\gamma}(s,a)|s \sim d^b] && (6.2)\\ &= E[\rho(s,a)\tau(s,a)Q^{\pi,\gamma}(s,a)|s \sim d^b, a \sim b(|s)] && (6.3)\\ &= E_b[\rho(s,a)\tau(s,a)Q^{\pi,\gamma}(s,a)] && (6.4) \end{aligned}$$

By the Policy Improvement Theorem for g(u):

Given any policy parameter u, let

$$u' = u + \alpha g(u)$$

Then there exists an $\epsilon > 0$ *such that for all positive* $\alpha < \epsilon$

$$J_\gamma(u') \geq J_\gamma(u)$$

and the Off-Policy Gradient Theorem, stating that:
Given $U \subset \mathbb{R}^{N_u}$ *a non-empty and compat set:*

$$\tilde{Z} = \{u \in U | g(u) = 0\} \tag{1}$$
$$Z = \{u \in U | \nabla_u J_\gamma(u) = 0\} \tag{2}$$

where $\tilde{Z}$ and Z are respectively the set of parameter vector at local minima obtained using the true gradient $\nabla_u J_\gamma(u)$ and approximate gradient g(u), we know that we can use g(u) as gradient to update parameters u instead of computing for the original policy gradient.

Sutton et al.2000 had proven that introducing baseline will not change the expected value of gradient, so here we subtract the Q with value provided by the critic:

$$g(u) = E_b[\rho(s,a)\psi(s,a)(Q^{\pi,\gamma}(s,a) - \hat{V}(s))] \tag{7}$$

The last step left is to futher approximate the action-value (with related to policy π) , by the off-policy λ return:

$$g(u) \approx \hat{g(u)} = E_b[\rho(s,a)\psi(s,a)(R_t^\lambda - \hat{V}(s))] \tag{8}$$

where the off-policy λ return is defined by:

$$R_t^\lambda = r_{t+1} + (1-\lambda)d(s_{t+1})\hat{V}(s_{t+1}) + \lambda d(s_{t+1})\rho(s_{t+1}, a_{t+1})R_{t+1}^\lambda$$

we can see that the value of next time point is represented with the last two terms with different weights, (1 - λ) and λ.

Finally, the forward view of Off-PAC is:

$$u_{t+1} = u_t + \alpha_{u,t}\rho(s_t,a_t)\psi(s_t,a_t)(R_t^\lambda - \hat{V}(s))] \tag{9}$$

This update equation can be futher simplify to relief the λ return to make it possible for implementation: it has been shown in the appendix that this holds:

$$E_b[\rho(s,a)\psi(s,a)(R_t^\lambda - \hat{V}(s))] = E_b[\delta_t e_t]$$

where the update for δ is:

$$\delta_t = r_{t+1} + d(s_{t+1})\hat{V}(s_{t+1}) - \hat{V}(s_t)$$

and the update for e is:

$$e_t = \rho(s_t,a_t)(\psi(s_t,a_t) + \lambda e_{t-1})$$

For the convergence analysis part, they first made the following assumptions:

- (A1) The policy viewed as a function of u, $\pi(a|s) : \mathbb{R}^{N_u} \to (0,1]$, is continuously differentiable $\forall s \in S, a \in A$
- (A2) The update on u_t includes a projection operator, Γ: $\mathbb{R}^{N_u} \to \mathbb{R}^{N_u}$, that projects any u to a compactset $U = \{u|q_i(u) \le 0, i = 1...s\} \subset \mathbb{R}^{N_u}$ where $q_i(\cdot) : \mathbb{R}^{N_u} \to \mathbb{R}$ are continuously differentiable functions specifying the constraints of a compact region. For u on the boundary of U, the gradient of the active q_i are linearly independent. Assume the compact region is large enought to contain at least one (local) maximum of J_γ.
- (A3) The behavior policy has a minimum positive value $b_{\min} \in (0,1] : b(a|s) \le b_{\min}, \forall s \in S, a \in A$
- (A4) The sequence $(x_t, x_{t+1}, r_{t+1})_{t \le 0}$ is i.i.d and has uniformly bounded second moments.
- (A5) For every $u \in U, V^{\pi,\gamma} : S \to R$ is bounded
- (P2) Matrices $C = E[x_t x_t^T], A = E[x_t(x_t - \gamma x_t)^T]$ are non-singular and uniformly bounded.
- (S1) $\alpha_{v,t} = \alpha_{w,t} = \alpha_{u,t} = \infty$ and $\sum_t(\alpha_{v,t}^2) < \infty, \sum_t(\alpha_{w,t}^2) < \infty, \sum_t(\alpha_{u,t}^2) < \infty$ with $\frac{\alpha_{u,t}}{\alpha_{v,t}} \to 0$
- (S2) Define $H(A) \doteq (A + A^T)/2$ and let $\lambda_{\min}(C^{-1}H(A))$ be the minimum eigenvalue of the matrix $C^{-1}H(A)$. Then $\alpha = \eta\alpha_{v,t}$ for some $\eta > \max(0, -\lambda_{\min}(C^{-1}H(A)))$

(A1) Provides the bedrock of using gradient ascent
(A2) Makes the proof of boundedness easier
(A3) Ensures that every action has non-zero probabiblity of being taken for all states
(S1) The learning rate assumptions where $\alpha_{v,t}, \alpha_{u,t}$ are repectively the learning rate of critic, actor at time t . $\alpha_{u,t}/\alpha_{v,t} \to 0$ is to make sure that critic is trained at a faster rate. (S2)

4 Conclusion

The paper compares three other different algorithms with the proposed method, which are:

1 $Q(\lambda)$ (Q-learning with $\lambda = 0$)
2 Greedy-GQ (GQ(λ) with greedy policy)
3 Softmax-GQ (GQ(λ) with softmax policy)

under three different openAI-gym environments (MountainCar, CartPole, Continous grid world). We see from the statistics that Off-PAC outperforms the other algorithms on all of the three environments.

This paper suggests that Off-PAC is more robust to the noise because it has lower variance than the action-value based methods.

Here the convergence property requires that $\lambda = 0$ but should be extend to $\lambda > 0$ in the future Futhermore, the statistics are retrieved using only a small set of possible hyperparameters, perhaps a test with more of them help stregthen the convergence property.

References

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

[Degris et al., 2012]