# RL Theory Project

July 14, 2020

Vu Thu Thao Huong - 0856156

Paper: Konda and Tsitsiklis, Actor-Critic Algorithms, NIPS 2000

## 1 Introduction

The majority of Reinforcement Learning methods are based on either Actor-only methods or Critic-only methods. Actor methods try to find the optimal policies by using algorithms like Policy Gradients while Critic methods attempt to find or approximate the optimal value function. Each method has its own advantages and disadvantages. As a result, the main purpose of this paper is to use a hybrid of Actor and Critic methods to take advantage of all the good while eliminating the drawbacks from the two methods.

The principal idea is to split the model in two: Actor which decides which action to take and Critic to evaluates how good was the action and how it should adjust. Those two models participate in a game where they both get better in their own role as the time passes. The result is that the overall architecture will learn to play the game more efficiently than the two methods separately.

In the algorithm provided by this paper the critic is estimated using TD learning in which the critic need not to compute the exact value function which are high-dimensional but instead will use linear approximation. Estimated value function will then be used to update the actor parameters. The paper has shown that the algorithm converges.

The paper provides one approach to Actor-Critic algorithm. There are other alternatives such as A2C/A3C or estimating the critic using REINFORCE algorithm without having to estimating the value function. The key observation for this paper is that actor and critic parameters should not be selected independently, instead the critic value function can be completely determined by the actor parameters. In addition, unlike other methods, the algorithm described in this paper can be applied to high dimensional problems.

# 2 Problem Formulation

The actor-critic algorithms is designed for simulation-based optimization of a Markov decision process over a parameterized family of randomized stationary policies.

Consider a Markov decision process with finite state space $S$, and finite action space $A$. A randomized stationary policy (RSP) is a mapping $\mu$ that assigns to each state $x$ a probability distribution over the action space $A$. We consider a set of randomized stationary policies $\mathbb{P} = \{\mu_\theta; \theta \in \mathbb{R}^n\}$, parameterized in terms of a vector $\theta$.

Assumptions about $\mathbb{P}$:

- (A1) For all $x \in S$ and $u \in A$ the map $\theta \mapsto \mu_\theta(x,u)$ is twice differentiable with bounded first, second derivatives. Furthermore, there exists a $\mathbb{R}^n$ valued function $\psi_\theta(x,u)$ such that $\nabla\mu_\theta(x,u) = \psi_\theta(x,u)\psi_\theta(x,u)$ where the mapping $\theta \mapsto \psi_\theta(x,u)$ is bounded and has first bounded derivatives for any fixed $x$ and $u$. Under (A1), if $\mu_\theta(x,u) \neq 0 : \psi_\theta(x,u) = \nabla \ln \mu_\theta(x,u)$
- (A2) For each $\theta \in \mathbb{R}^n$ , the Markov chains $\{X_n\}$ and $\{X_n, U_n\}$ are irreducible and aperiodic, with stationary probabilities $\pi_\theta(x)$ and $\eta_\theta(x,u) = \pi_\theta(x)\mu_\theta(x,u)$ , respectively, under the RSP $\mu_\theta$.

We have some following notations:

- $g : S \times A \to \mathbb{R}$: cost function
- $\mu_\theta(x,u)$ : probability of taking action $u$ when the state $x$ is encountered, under the policy corresponding to $\theta$.
- $p_{xy}(u)$ : probability that the next state is $y$, given that the current state is $x$ and the current action is $u$.

# 3 Theoretical Analysis

## 3.1 Cost function

Let $\lambda_\theta(x,u) = \sum_{x \in S, u \in A} g(x,u)\eta_\theta(x,u)$ be the avearge cost function. Our goal is to minize $\lambda(\theta)$ over $\theta$. Let $V_\theta : S \mapsto \mathbb{R}$ :"differential" cost function or the expected excess cost on top of the average cost at state $x$, defined as solution to the Poisson equation:

$$\lambda(\theta) + V_\theta(x) = \sum_{u \in A} \mu_\theta(x,u) \left[ g(x,u) + \sum_y p_{xy}(u) V_\theta(y) \right]$$

**Theorem 1.** The gradient for optimization problem can be calculated as:

$$\nabla\lambda(\theta) = \sum_{x,u} \eta_\theta(x,u) q_\theta(x,u) \psi_\theta^i(x,u) \tag{1}$$

Define q-function $q_\theta$ : $S \times A \to \mathbb{R}$ : expected excess cost incurred over a certain renewal period of the Markov chain $\{X_n, U_n\}$, under the RSP $\mu_\theta$, and is then estimated by means of simulation leading to actor-only algorithms:

$$q_\theta(x,u) = g(x,u) - \lambda(\theta) + \sum_y p_{xy}(u) V_\theta(y)$$

where and $\psi_\theta^i(x,u)$ is the ith component of $\psi_\theta$

For any $\theta \in \mathbb{R}^n$ , we define the inner product $\langle q_1, q_2 \rangle$ of two real valued functions $q_1$ , $q_2$ on $S \times A$, viewed as vectors in $\mathbb{R}^{|S||A|}$, by

$$\langle q_1, q_2 \rangle_\theta = \sum_{x,u} \eta_\theta(x,u) q_1(x,u) q_2(x,u)$$

Formula (1) can be rewritten as:

$$\nabla\lambda(\theta) = \langle q_\theta, \psi_\theta^i \rangle_\theta$$

Let $||.||_\theta$ denote the norm induced by this inner product on $\mathbb{R}^{|S||A|}$. For each $\theta \in \mathbb{R}^n\}$ let $\Psi_\theta$ denote the span of the vectors $\{\psi_\theta^i; 1 \le i \le n\}$ in $\mathbb{R}^{|S||A|}$. Although the gradient of $\lambda$ depends on the q-function, which is a vector in a possibly very high dimensional space $\mathbb{R}^{|S||A|}$, the dependence is only through its inner products with vectors in $\Psi_\theta$ . Thus, instead of "learning" the function $q_\theta$,it would suffice to learn the projection of $q_\theta$ on the subspace $\Psi_\theta$. Therefore, let $\Pi_\theta : \mathbb{R}^{|S||A|} \mapsto \Psi_\theta$ be the projection operator defined by: $\Pi_\theta q = arg\ min_{\hat{q} \in \Psi_\theta} ||q - \hat{q}||_\theta$. Since $\langle q_\theta, \psi_\theta \rangle_\theta = \langle \Pi_\theta q, \psi_\theta \rangle_\theta$ (2), it is enough to compute the projection of $q_\theta$ onto $\Psi_\theta$ to compute $\nabla\lambda$.

*Proof of Theorem 1:* If we have $q$ is solution of the Poisson equation with parameter $\theta$, we can rewrite the q-function as:

$$q_\theta = g - \lambda(\theta)\underline{1} + p_\theta q_\theta$$

Differentiate both sides of equation w.r.p to $\theta$:

$$\nabla\lambda(\theta)\underline{1} + \nabla q_\theta = p_\theta(\psi_\theta q_\theta) + p_\theta(\nabla q_\theta)$$

Taking inner product with $\underline{1}$ on both sides of the equation and using the following facts: $\langle 1, p_\theta f \rangle_\theta = \langle 1, f \rangle_\theta$ and $\langle 1, \psi_\theta \rangle_\theta = 0$. We can prove:

$$\nabla\lambda(\theta) = \langle q_\theta, \psi_\theta \rangle_\theta$$

### 3.2 Actor-Critic Algorithms

In this paper, the actor critic-algorithms is considered as stochastic gradient algorithms on the parameter space of the actor. When the actor parameter vector is $\theta$, the job of the critic is to compute an approximation of the projection of $\Pi_\theta q$ $q_\theta$ onto $\Psi_\theta$ . Two actor-critic algorithms are provided is this paper which differs only on how the critic updates are concerned. In both variants, the critic is a TD algorithm with a linearly parameterized approximation architecture for the q-function:

$$Q_r^\theta(x,u) = \sum_{j=1}^{m} r^j \phi_\theta^j(x,u)$$

where $r$ denotes the $m$-parameter vector of the critic and $\phi_\theta^j$ are features used by the critic are dependent on the actor parameter To incorporate described cost function making sure equation (2) hold we would let:

- $\Pi_\theta$ is redefined as projection onto $\Phi_\theta$ as long as $\Phi_\theta$ contains $\Psi_\theta$
- $m = n$ and $\phi_\theta^i = \psi_\theta^i$

**Update for the critic:**

$$\lambda_{k+1} = \lambda_k + \gamma_k(g(X_k,U_k) - \lambda_k)$$

$$r_{k+1} = r_k + \gamma_k \left(g(X_k,U_k) - \lambda_k + Q_{r_k}^{\theta_k}(X_{k+1},U_{k+1}) - Q_{r_k}^{\theta_k}(X_k,U_k)\right) z_k$$

where $\gamma_k$ is the positive step size parameter.

We can use either $TD(1)$ or $TD(\alpha)$ to update $z_k$:

– $TD(1)$ *Critic*: Let $x^*$ be a state in $S$:
$z_{k+1} = z_k + \phi_{\theta_k}(X_{k+1},U_{k+1})$ *if* $X_{k+1} \neq x^*$
*otherwise* $z_{k+1} = \phi_{\theta_k}(X_{k+1},U_{k+1})$

– $TD(\alpha)$ *Critic*: $0 \leq \alpha \leq 1$:

$$z_{k+1} = \alpha z_k + \phi_{\theta_k}(X_{k+1},U_{k+1})$$

**Update for the Actor:**

$$\theta_{k+1} = \theta_k - \beta_k \Gamma(r_k) Q_{r_k}^{\theta_k}(X_{k+1},U_{k+1}) \psi_{\theta_k}(X_{k+1},U_{k+1})$$

where $\beta_k$ is a positive stepsize and $\Gamma(r_k)$ is a nomalization factor satisfying additional assumptions:

– (A3) $\Gamma(.)$ is Lipschitz continuous.
– (A4) There exists $C > 0$ such that: $\Gamma(r) \leq \frac{C}{1+||r||}$

### 3.3 Convergence of actor-critic algorithms

Some further assumptions are needed:

- (A5) For each $\theta \in \mathbb{R}^n$, we define an $m \times m$ matrix $G(\theta)$ by: $G(\theta) = \sum_{x,u} \eta_\theta(x,u)\phi_\theta(x,u)\phi_\theta(x,u)^T$. $G(\theta)$ is positive definite, that is, there exists some $\varepsilon_1 > 0$ such that for all $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$: $r^T G(\theta) r \geq \varepsilon_1 ||r^2||$
- (A6) We assume that the step size sequences $\{gamma_k\}$, $\{beta_k\}$ are positive, non-increasing, and satisfy: $\delta_k > 0 \; \forall k$, $\sum_k \delta_k = \infty$, $\sum_k \delta_k^2 < \infty$, $\frac{\beta_k}{\gamma_k} \to 0$ where $\delta_k$ stands for either $\beta_k$ or $\gamma_k$.

As the paper used gradient-based actor-critic algorithms, convergence to a globally optimal policy (within the given class of RSP's) cannot be expected, but instead the algorithm can have local minimum of $\lambda(\theta)$ or in other words $\nabla\lambda(\theta)$ should converge to 0.

**Theorem 2.** In an actor-critic algorithm with a TD(1) critic: $\liminf_k ||\nabla\lambda(\theta)|| = 0 \;\; w.p.1$

**Theorem 3.** For any $\varepsilon > 0$, there exists some $\lambda$ close to 1, so that the algorithm that uses a $TD(\alpha)$ critic satisfies $\liminf_k ||\nabla\lambda(\theta)|| \leq \varepsilon \;\; w.p.1$

*Proof of Theorem 2 and 3 :* The proof will follow section 6 of the paper "On actor-critic algorithms" from the same author.

For every $\theta \in \mathbb{R}^n$, let: $H_\theta = \psi_\theta(x,u)\phi'_\theta(x,u)$ and $\overline{H} = \langle \psi_\theta, \phi'_\theta \rangle_\theta$.

Let $\overline{r}(\theta)$ be such that $\overline{h}_1(\theta) = \overline{G}_1(\theta)\overline{r}(\theta)$, so that $\overline{r}(\theta)$ is the limit of the critic parameter $r$ if the policy parameter $\theta$ was held fixed. The recursion for the actor parameter $\theta$ can be written as:

$\theta_{k+1} = \theta_k - \beta_k H\theta_k(X_{k+1}, U_{k+1}(r_k\Gamma(r_k)))$

$\theta_{k+1} = \theta_k - \beta_k \overline{H}(\theta_k)(\overline{r}(\theta_k)\Gamma(\overline{r}(\theta_k))) - \beta_k(H\theta_k(X_{k+1}, U_{k+1}) - \overline{H}(\theta_k))(r_k\Gamma(r_k)) - \beta_k\overline{H}(\theta_k)(r_k\Gamma(r_k) - \overline{r}(\theta_k)\Gamma(\overline{r}(\theta_k)))$

Let:

$f(\theta) = \overline{H}(\theta)\overline{r}(\theta)$

$e_k^{(1)} = (H\theta_k(X_{k+1}, U_{k+1}) - \overline{H}(\theta_k))(r_k\Gamma(r_k))$

$e_k^{(2)} = \overline{H}(\theta_k)(r_k\Gamma(r_k) - \overline{r}(\theta_k)\Gamma(\overline{r}(\theta_k)))$

Using Taylor's series expansion:

$\lambda(\theta_{k+1}) \leq \lambda(\theta_k) - \beta_\theta\Gamma(\overline{r}(\theta_k))$
$\nabla\lambda(\theta)f(\theta_k) - \beta_\theta\nabla\lambda(\theta)e_k^{(1)} - \beta_\theta\nabla\lambda(\theta)e_k^{(2)} + C\beta_\theta^2|H\theta_k(X_{k+1}, U_{k+1})(r_k\Gamma(r_k))|^2$
(3)

Fix some $T > 0$ and define a sequence $k_j$ by:

$k_0 = 0$ and $k_{j+1} = min\{k \geq k_j | \sum_{l=k_j}^{k} \beta_k \geq T\} \;\; for\; j > 0$

Using (3) we have:

$$\lambda(\theta_{k+1}) \leq \lambda(\theta_k) - \sum_{k=k_j}^{k_{j+1}-1} \beta_k(|\nabla\lambda(\theta)|^2 - C(1-\lambda)|\nabla\lambda(\theta)|) + \delta_j$$

where $\delta_j$ is defined as:

$$\delta_j = \sum_{k=k_j}^{k_{j+1}-1} (\beta_k \nabla\lambda(\theta)(e_k^{(1)} + e_k^{(2)}) + C\beta_\theta^2 |H\theta_k(X_{k+1}, U_{k+1})(r_k\Gamma(r_k))|^2)$$

It can be proved that $\delta_j$ goes to zero. If the result fails to hold, it can be shown that if the sequence $\lambda(\theta_k)$ would decrease indefinitely, it would contradict the boundedness of $\lambda(\theta_k)$. So Theorem 2 and 3 shall hold.

# 4 Conclusion

The paper has provided one approach to Actor-Critic Algorithms which can be applied to high-dimensional problems. The algorithm shows desired convergence property under certain assumptions. Further research can focus on how relaxing mentioned assumptions affecting the algorithms results. We can also empirically evaluate how such Actor-Critic Algorithms outperform Actor-only or Critic - only methods.