Actor-Critic Algorithms

Chen, Bo-Jen
Institute of Finance
National Chiao Tung University
bb19x11@gmail.com

1 Introduction

Please provide a clear overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper
- The high-level technical insights into the problem of interest
- The main contributions of the paper (compared to the prior works)
- Your personal perspective on the proposed method

This paper proposes some actor-critic algorithms and provide an overview of a convergence proof. They show the critic should ideally compute a certain projection of the value function onto a low-dimensional subspace spanned by a set of basis functions determined completely by the parameterization of the actor. In other words, the actor parameterization and the critic parameterization are not supposed to be chosen independently. In my opinion, this paper make a "real" connection between the actor and critic, which truly combines the advantages of actor-only and critic-only methods.

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

- Your notations (e.g. MDPs, value functions, function approximators,...etc)
- The optimization problem of interest
- The technical assumptions

Consider a MDP with finite space $\mathcal{S}$, and finite action space $\mathcal{A}$. g : $\mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is a given cost function. A randomized stationary policy (RSP) is a mapping μ assigning to each state x a probability distribution over the action space. A set of randomized stationary policy $P = \{ \mu_\theta \,|\, \theta \in R^n \}$, parameterized in terms of vector θ. For each pair $(x, \mu) \in \mathcal{S} \times \mathcal{A}$, μ denotes the probability of taking action μ when the state x is encountered, under the policy corresponding to θ.$p_{xy}(\mu)$ denotes the probability that the next state is y, given the current states is x and the current action is μ.

The authors make the following assumption (A1) and (A2) :

(A1) $\forall x \in \mathcal{S}, \mu \in \mathcal{A}$, the map $\theta \mapsto \mu_\theta(x, \mu)$ is twice differentiable with bounded first and second derivatives. Moreover, there exists a $\mathcal{R}^n$-valued function $\psi_\theta(x, \mu)$ such that $\bigtriangledown \mu_\theta(x, \mu) = \mu_\theta(x, \mu)\psi_\theta(x, \mu)$ where the mapping $\theta \mapsto \mu_\theta(x, \mu)$ is bounded and has first bounded derivatives for any fixed x and μ.

Preprint. Under review.

For each $\theta \in \mathcal{R}^n$, the Markov chains $\{\mathcal{X}_n\}$ and $\{\mathcal{X}_n, \mathcal{U}_n\}$ are irreducible and aperiodic, with stationary probabilities $\pi_{\theta(x)}$ and $\eta_\theta(x,\mu) = \pi_\theta(x,\mu)$ respectively, under the RSP. $\{\S_n\}$ is the sequence of states and $\{\mathcal{U}_n\}$ is the sequence of actions.

Consider the average cost function :
$\lambda(\theta) : \mathcal{R}^n \mapsto \mathcal{R}$, given by $\lambda(\theta) = \sum_{x\in\S,\mu\in\mathcal{A}} g(x,\mu)\eta_\theta(x,\mu)$. The authors are interested in minimizing $\lambda(\theta)$ overall θ.

For each $\theta \in \mathcal{R}^n$, let $\mathcal{V}_\theta : \mathcal{S} \mapsto \mathcal{R}$ be the differential cost function. $\lambda(\theta) + \mathcal{V}_\theta = \sum_{\mu\in\mathcal{A}} \mu_\theta(x,\mu)[g(x,\mu) + \sum_y p_{xy}(\mu)\mathcal{V}_\theta(y)]$

The authors define $< q_1, q_2 >_\theta = \sum_{x,\mu} \eta_\theta(x,\mu)q_1(x,\mu)q_2(x,\mu)$ and therefore rewrite $\frac{\partial\lambda(\theta)}{\partial\theta}$ as $< q_\theta, \psi^i_\theta >, i = 1\ n$. For each $\theta \in \mathcal{R}^n$, Ψ_θ denote the span of the valued function vectors $\{\psi^i_\theta : i = 1\ n\}$ in $R^{|\mathcal{S}||\mathcal{A}|}$

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight your new proofs or extensions (if any).

The critic is a TD with linearly parameterized approximation architecture for the q-function, of the form $\mathcal{Q}^\theta_r = \sum_{j=1}^m r^j\varphi^j_\theta(x,\mu)$ where $r = (r^i, ..., r^m)$ denote the parameter vector of critic. The features $\varphi^j_\theta, j = 1\ m$, used by the critic are dependent on the actor parameter vector θ and are chosen such that their span in $\mathcal{R}^{|\mathcal{S}||\mathcal{A}|}$, denoted by Φ_θ, contains Ψ_θ. The authors allow the possibility that $m > n$ and Φ_θ properly contains Ψ_θ. In other words, the critic uses more feature than that are actually necessary. This decision makes the set of features richer, which avoid ill-conditioned projection and make the algorithms more stable. Besides, the use of additional features can result in a reduction of the approximation error for $TD(\alpha)$ proposed by this paper (where $\alpha < 1$).

Let r_k, z_k, λ_k be the parameters of critic, where r is the parameter vector, λ is the estimate of the average cost, z s an m-vector that represents Sutton's eligibility trace, at time k. The critic carries out an update in the following way :
$\lambda_{k+1} = \lambda_k + \gamma_k(g(\mathcal{X}_k,\mathcal{U}_k) - \lambda_k)$
$r_{k+1} = r_k + \gamma_k(g(\mathcal{X}_k,\mathcal{U}_k) - \lambda_k + \mathcal{Q}^{\theta_k}_{r_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1}) - \mathcal{Q}^{\theta_k}_{r_k}(\mathcal{X}_k,\mathcal{U}_k))$
where γ_k is a positive step size parameter.

$TD(1)$ critic :
Let $x^\star$ be a state in $\mathcal{S}$.
$z_{k+1} = z_k + \phi_{\theta_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1})$ if $\mathcal{X}_{k+1} \neq x^\star$
$z_{k+1} = \phi_{\theta_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1})$, otherwise.

$TD(\alpha)$ critic with $0 \leq \alpha < 1$ is of the form : $z_{k+1} = \alpha z_k + \phi_{\theta_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1})$

As for the actor, $\theta_{k+1} = \theta_k - \beta_k\Gamma(r_k)\mathcal{Q}^{\theta_k}_{r_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1})\psi_{\theta_k}(\mathcal{X}_{k+1},\mathcal{U}_{k+1})$
where β_k is a positive step size parameter.

In the algorithm mentioned above, the observations from all past epsiodes affect current critic parameter r. Since as long as θ is changing slowly, the observations from recent epsiodes carry useful information on q-function under the current policy, in this sense critic is learning. Hence it could be advantageous.

Since the algorithm proposed by this paper is gradient-based, one can not expect to prove convergence to a globally optimal policy. Therefore, one could hope for is the convergence of gradient of average cost function λ to zero. To get this result, another assumption is introduced.

Assume the step size sequences $\{\gamma_k\}$ and $\{\beta_k\}$ are positive, non-increasing, and satisfy:

$\forall k$, $\delta_k > 0$, $\sum_k \delta_k = \infty, \sum_k \delta_k^2 < \infty$ where δ stands for γ_k and β_k. This paper assume that $\frac{\gamma_k}{\beta_k} \to 0$. Next, there is one theorem based on the assumption.

Theorem :
In an actor-critic algorithm with a $TD(1)$ critic, $\liminf_k || \bigtriangledown \lambda_k(\theta_k)||$ with probability 1. Furthermore, if $\{\theta_k\}$ is bounded with probability 1, then $\lim_k || \bigtriangledown \lambda_k(\theta_k)|| = 0$ with probability 1.

Since $\frac{\gamma_k}{\beta_k}$ approaches zero, the size of the actor updates becomes negligible compared to the size of the critic updates. Therefore the actor looks stationary, as far as the critic is concerned.

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations
- Any latest results on the problem of interest

Since the number of parameters that the actor has to update is relatively small compared with the number of states, it is not useful to have the critic attempting to compute the exact value function, which is also a high-dimensional object. Instead, it should compute a projection of the value function onto a low-dimensional subspace spanned by a set of basis functions, which are completely determined by the parameterization of the actor.

References