On the Finite-Time Convergence of Actor-Critic Algorithm

Shuang Qiu
University of Michigan

Zhuoran Yang
Princeton University

Jieping Ye
University of Michigan

Zhaoran Wang
Northwestern University

Abstract

Actor-critic algorithm and their extensions have made great achievements in real-world decision-making problems. In contrast to its empirical success, theoretical understanding of actor-critic seems unsatisfactory. Most existing results only show asymptotic convergence, which were developed mainly based on approximating the dynamic system of the actor and critic using ordinary differential equations. However, the finite-time convergence analysis of actor-critic algorithm remains to be explored. The main challenges lie in the nonconvexity of parameterized policies, the coupling of the updates for actor and critic, and the data sampling dependency in online settings. In this paper, we provide a finite-time convergence analysis for an online actor-critic algorithm under the infinite-horizon average reward setting. In critic step, we give a theoretical analysis of TD(0) algorithm for the average reward with dependent data in online settings. Besides, we show that the sequence of actor iterates converges in sublinear rate to a stationary point up to some irremovable bias due to the value function approximation by linear functions. To the best of our knowledge, our work seems to provide the first finite-time convergence analysis for online actor-critic algorithm under the infinite-horizon average reward setting.

1 Introduction

When combined with deep neural networks [Goodfellow et al., 2016], actor-critic algorithm [Konda and Tsitsiklis, 2000] and their extensions [Schulman et al., 2015, 2017] have made astonishing achievements in complicated real-world decision-making problems [Silver et al., 2016, 2017, OpenAI, 2018, Vinyals et al., 2019]. Here the actor corresponds to the policy and the critic estimates the value function associated with the actor. In each iteration of the algorithm, the actor performs policy gradient [Sutton et al., 2000] updates using the value function provided by the critic, and the critic simultaneously updates its parameter using policy evaluation algorithms such as temporal difference (TD) learning [Sutton, 1988]. Thus, actor-critic inherit the advantages of both policy- and value-based reinforcement learning methods – it can be readily applicable to continuous action spaces [Silver et al., 2014, Lillicrap et al., 2016] and enjoys small variance by incorporating TD learning.

In contrast to the huge empirical success, theoretical understanding of actor-critic seems unsatisfactory, with most existing results showing asymptotic convergence with linear value functions. These proofs mainly depend on approximating the dynamic system of the actor and critic using ordinary differential equations (ODE) [Borkar, 2008, Borkar and Meyn, 2000, Kushner and Yin, 2003]. Such a distinct separation between theory and practice exists mainly due to the following two challenges. First, policy optimization for the family of parametrized policies is in general nonconvex. Thus, even finding a local minimum via gradient-based updates can be NP-hard in the worst case [Murty and Kabadi, 1987]. Second, the actor and critic have different optimization objectives and their updates

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

are coupled together, which makes it challenging to analyze their dynamics. Moreover, as we will see in §3, the optimization problems faced by the actor and critic can be unified into the framework of bilevel optimization, which is a class of NP-hard problems [Hansen et al., 1992] with mostly heuristic solution strategies [Colson et al., 2007, Sinha et al., 2017]. Third, in online reinforcement learning settings, data is no longer sampled independently, which brings further bias the learning process.

Despite these difficulties, under the infinite-horizon average reward setting, we study the finite-time convergence of online actor-critic algorithm with linear function approximation. In contrast to the existing asymptotic analysis based on ODE approximation, we decouple the actor and critic updates by appropriately choosing the stepsizes such that their analysis can be carried out alternatively. We establish finite-time convergence of the TD(0) used by the critic with a trajectory of dependent data. More importantly, due to the error incurred by the critic, we show that the actor essentially performs policy gradient updates with bias and also establish its convergence rate in expectation.

Main Contribution. Our contribution is listed as follows: **(1)** After decoupling the actor and critic updates, for policy evaluation, we provide a finite-time convergence analysis for TD(0) algorithm with dependent data under the average reward setting. **(2)** We show that the sequence of actor iterates converge in sublinear rate to a stationary point up to some irremovable bias due to approximating the value function using linear functions. **(3)** We seem to provide the first finite-time convergence analysis for online actor-critic algorithm under the infinite-horizon average reward setting. Our analysis framework is general and can be extended to other variations of actor-critic algorithms.

2 Background

The infinite-horizon average reward reinforcement learning problem is modeled as an average reward Markov Decision Process (MDP). Suppose that $\mathcal{S}$ and $\mathcal{A}$ are the finite state space and finite action space respectively. The policy π is defined as a function that $\pi : \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ such that $\pi(a \,|\, s)$ is the probability of choosing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$. From a practical perspective, the policy π is usually parameterized by $\theta \in \Theta$ in a nonconvex form and then we denote the parameterized policy by π_θ. An agent takes action $a \sim \pi_\theta(\cdot \,|\, s)$ at state s. Letting $\mathbb{P}(s' \,|\, a, s)$ be the probability of an agent moving from state s to state s' with an action a, we can have a Markov transition probability induced by π_θ as $\mathbb{P}^{\pi_\theta}(s' \,|\, s) = \sum_{a \in \mathcal{A}} \mathbb{P}(s' \,|\, a, s)\pi_\theta(a \,|\, s)$ be the probability of moving from state s to state s'. At each time τ, we use a tuple $(s_\tau, a_\tau, s'_\tau, r_{\tau+1})$ to denote that an agent at state s_τ chooses an action a_t and transit to a new state s'_τ with a reward $r_{\tau+1} := r(s_\tau, a_\tau, s'_\tau)$, where $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is a reward mapping. We make the following assumption on the policy π_θ and the probability $\mathbb{P}(s' \,|\, a, s)$.

Assumption 2.1. *The policy π_θ and $\mathbb{P}(s' \,|\, a, s)$ guarantee that the Markov chain decided by $\mathbb{P}^{\pi_\theta}(s' \,|\, s)$ for any $\theta \in \Theta$ is irreducible and aperiodic. Thus, there always exists a unique stationary distribution for any $\theta \in \Theta$, which is denoted by $d^{\pi_\theta}(s), \forall s \in \mathcal{S}$.*

Assumption 2.1 is a common assumption for actor-critic algorithms [Konda and Tsitsiklis, 2000, Bhatnagar et al., 2009]. It essentially implies that the state transition probability $\mathbb{P}^{\pi_\theta}(s' \,|\, s)$ with any parameter $\theta \in \Theta$ is associated with irreducible aperiodic Markov chain.

The goal of the agent is to find a policy π_θ such that the nonconvex infinite-horizon average reward function $J(\pi_\theta)$ is maximized:

$$\underset{\theta \in \Theta}{\text{maximize}} \quad J(\pi_\theta) = \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_{t+1} \,\middle|\, \pi_\theta\right] = \mathbb{E}_{(s,a,s') \sim \Upsilon_\theta}[r(s, a, s')], \tag{1}$$

where $\mathbb{E}_{(s,a,s') \sim \Upsilon_\theta}$ is short for $\mathbb{E}_{s \sim d^{\pi_\theta}(\cdot), a \sim \pi_\theta(\cdot \,|\, s), s' \sim \mathbb{P}(\cdot \,|\, s, a)}$ and Υ_θ is a joint distribution of s, a, s'. Specifically, in this paper, we consider the case that $\Theta = \mathbb{R}^m$ where m is the dimension of θ.

The action-value function Q^{π_θ} and state-value function V^{π_θ} are defined as

$$Q^{\pi_\theta}(s, a) = \lim_{T \to \infty} \sum_{t=0}^{T-1} \mathbb{E}[r_{t+1} - J(\pi_\theta) \,|\, s_0 = s, a_0 = a, \pi_\theta], \quad V^{\pi_\theta}(s) = \mathbb{E}_{a \sim \pi_\theta(\cdot \,|\, s)}[Q^{\pi_\theta}(s, a)].$$

3 Actor-Critic Algorithm

In this section, we first present the policy gradient theorem, which plays a critical role in updating θ via policy gradient. Then, we show that the actor-critic algorithm essentially solves a certain bilevel

optimization problem. Following this point of view, we further present the actor step and the critic step respectively in detail.

Theorem 3.1 (Policy Gradient Theorem)**.** *The gradient of $J(\pi_\theta)$ can be written as follows*

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}(\cdot), a\sim\pi_\theta(\cdot\,|\,s), s'\sim\mathbb{P}(\cdot\,|\,a,s)}[\delta^{\pi_\theta} \cdot \nabla \log \pi_\theta(a\,|\,s)], \tag{2}$$

where we define the temporal difference (TD) error by $\delta^{\pi_\theta} := r(s,a,s') - J(\pi_\theta) + V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$.

Please refer to Supplementary Material §B for detailed proof which is based on Sutton et al. [2000].

To approximate V^{π_θ} in Theorem 3.1, one typical approach is to parameterize the state-value function by a parameter $\omega \in \Omega$. Here we denote by V_ω the parameterized state-value function, which can be in a linear form that for all $s \in \mathcal{S}$,

$$V_\omega(s) = \phi(s)^\top \omega, \tag{3}$$

where $\omega \in \mathbb{R}^n$ and $\phi : \mathcal{S} \mapsto \mathbb{R}^n$ is a basis function to generate n-dimensional features for any state s.

Given the above settings, we unify the optimization problem faced by the actor and critic from the perspective of bilevel optimization in the following proposition.

Proposition 3.2. *The actor-critic algorithm for average reward reinforcement learning can be interpreted as a bilevel optimization problem in the following form*

$$\underset{\theta\in\mathbb{R}^m}{\text{maximize}} \quad J(\pi_\theta) = \mu^*(\theta) \tag{4}$$

$$\text{subject to} \quad \{\omega^*(\theta), \mu^*(\theta)\} = \underset{\omega\in\Omega, \mu\in\mathbb{R}}{\operatorname{argmin}} \; h(\omega, \mu; \theta), \tag{5}$$

Specifically, the function $h(\omega, \mu; \theta)$ is defined as

$$h(\omega,\mu;\theta) = \left\|\mathbb{E}_{(s,a,s')\sim\Upsilon_\theta}\{[r - \mu + V'_\omega - V_\omega]\phi\}\right\|^2_{C_\theta^{-1}} + \left\|\mathbb{E}_{(s,a,s')\sim\Upsilon_\theta}(r) - \mu\right\|_2^2,$$

with $C_\theta = \mathbb{E}_{s\sim d^{\pi_\theta}}(\phi\phi^\top)$, $r = r(s,a,s')$, $\phi = \phi(s)$, $V'_\omega = V_\omega(s')$, and $V_\omega = V_\omega(s)$ for abbreviation.

In this proposition, the upper-level problem (4) aims to maximize the infinite-horizon average reward $J(\pi_\theta)$, while the lower-level problem (5) is associated with policy evaluation, whose objective is close to mean-squared Projected Bellman error (MSPBE) [Maei, 2011]. Note that C_θ is an invertible matrix, which can be guaranteed by our Assumption 2.1. As we can see, the upper-level and lower-level problems are coupled with each other. The upper-level problem aims to find maximal value of $\mu^*(\theta)$, which is a function of θ as well as the minimizer to lower-level problem under policy π_θ.

However, bilevel programming is NP-hard [Hansen et al., 1992]. Thus, we can rely on approximating the solution instead of an exact one. The actor-critic algorithm is an online stochastic approximation scheme to alternately solve the above bilevel optimization problem. Specifically, the actor step performs an online gradient ascent to update θ once given current ω and μ, which is associated with the upper-level optimization problem. The critic step is to estimate ω and μ by policy evaluation algorithms once given current π_θ, which is associated with the lower level optimization problem. With given π_θ, TD(0) [Sutton, 1988] is widely-adopted to solve the minimization problem in the lower level. In this paper, our analysis of critic step is focused on TD(0) algorithm with linear state-value function approximation under the infinite-horizon average reward setting.

The actor-critic algorithm is summarized in Algorithm 1 with TD(0) algorithm for critic step demonstrated in Algorithm 2. As shown in Algorithm 1, at step t, the critic step generate an inexact solution $\{\widehat{\mu}_t, \widehat{\omega}_t\}$ as an estimation of the true solution $\{\mu_t^* := \mu^*(\theta_t), \omega_t^* := \omega^*(\theta_t)\}$ to lower-level problem under the current policy π_{θ_t}, and then the actor step updates θ with $\{\widehat{\mu}_t, \widehat{\omega}_t\}$.

3.1 Actor Step

The actor step aims to update the parameter θ of the policy π_θ via solving upper level problem (4) by gradient ascent. Thus, it requires to approximate the policy gradient, i.e., $\nabla_\theta J(\pi_\theta)$.

The policy gradient theorem (Theorem 3.1) provides the formulation of $\nabla_\theta J(\pi_\theta)$. As shown in Equation (2), the true gradient of $J(\pi_\theta)$ is an expectation over all samples on the stationary distribution d^{π_θ} and policy π_θ under the parameter θ. In online setting, the data sample (s, a, s', r) is usually

Algorithm 1 Actor-Critic Algorithm for Infinite-Horizon Average Reward

1: **Input:** Parameterized policy π_θ.
2: **Initialize:** Randomly initialize θ_0 .
3: **for** $t = 0, ..., T-1$ **do**
4: **Critic Step:** Obtain $(\widehat{\omega}_t, \widehat{\mu}_t)$ as an estimation of (ω_t^*, μ_t^*) by policy evaluation algorithm.
5: **Actor Step:** Restart sampling (s_t, a_t, s_t') and update the policy parameter θ as follows

$$\delta_t = r_{t+1} - \widehat{\mu}_t + {\phi_t'}^\top \widehat{\omega}_t - \phi_t^\top \widehat{\omega}_t,$$
$$\theta_{t+1} = \theta_t + \beta_{\theta,t} \cdot \delta_t \cdot \nabla_\theta \log \pi_{\theta_t}(a_t \mid s_t).$$

6: **end for**
7: **Return:** θ_T, $\widehat{\omega}_T$, $\widehat{\mu}_T$.

observed one by one. Thus, we need to make a stochastic approximation of the true gradient with one sample (s, a, s', r) in each step, which computed by $\delta_t \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)$ in line 5 of Algorithm 1.

However, this stochastic gradient is biased mainly due to: **(1)** the estimation error of the average reward $\widehat{\mu}_t^* = J(\pi_{\theta_t})$ by $\widehat{\mu}_t$; **(2)** the estimation error of the parameter ω_t^* by $\widehat{\omega}_t$; **(3)** the approximation error of the true state value $V^{\pi_{\theta_t}}$ by linear parameterization. In the section of theoretical analysis, we characterize these biases and reveal how they affect the convergence of the actor-critic algorithm.

We assume that for all $\theta \in \Theta$, there exists $\mathcal{E}_{\text{approx}} \geq 0$ such that

$$\|V^{\pi_\theta} - \Phi\omega^*(\theta)\|_2^2 \leq \mathcal{E}_{\text{approx}},$$

where we define $\Phi := [\phi(s^{(1)}), \phi(s^{(2)}), ...]^\top \in \mathbb{R}^{|\mathcal{S}| \times n}$ and $s^{(i)}$ represents the i-th state, and $\omega^*(\theta)$ is the optimal solution to problem (5) under policy π_θ. Note that $\mathcal{E}_{\text{approx}}$ can be an arbitrary non-negative scalar. Smaller $\mathcal{E}_{\text{approx}}$ indicates better approximation by the linear function.

In the actor step of Algorithm 1, the restarting sampling operation exists only for theoretical analysis to decouple the dependency of data sampled under different policy π_{θ_t}. In practice, this is not necessary since the dependency will be quite little if the iteration number is large in the policy evaluation algorithm.

3.2 Critic Step

The critic step can provide stochastic approximation to the solution of the lower-level problem in Equation (5) once given a policy π_θ. Specifically, at time t of Algorithm 1, given current θ_t, the critic step aims to solve the following problem

$$\underset{\omega \in \Omega, \mu \in \mathbb{R}}{\text{minimize}} \left\|\mathbb{E}_{(s,a,s') \sim \Upsilon_{\theta_t}}\{[r - \mu + V_\omega' - V_\omega]\phi(s)]\}\right\|_{C_{\theta_t}^{-1}}^2 + \left\|\mathbb{E}_{(s,a,s') \sim \Upsilon_{\theta_t}}(r) - \mu\right\|_2^2, \quad (6)$$

where Υ_{θ_t} is the joint distribution of s, a, and s' under the policy π_{θ_t}, and $C_{\theta_t} := \mathbb{E}_{s \sim \pi_{\theta_t}}[\phi\phi^\top]$. To solve the problem (6), we apply TD(0) algorithm shown in Algorithm 2. As TD(0) is an online algorithm, in each iteration, we only sample one data for updating. Note that we do not need to make an assumption of independent data sampling for TD(0) algorithm.

To understand the existence and uniqueness of the solution to Equation (6), similar to Tsitsiklis and Van Roy [1999], we make the following assumptions

Assumption 3.3. *The basis function $\phi : \mathcal{S} \mapsto \mathbb{R}^n$ where $\|\phi(\cdot)\|_2 \leq 1$ satisfies the following two assumptions: **(1)** The basis functions $\phi(s)$ for all $s \in \mathcal{S}$ are linearly independent, i.e., $\Phi = [\phi(s^{(1)}), \phi(s^{(2)}), ...]^\top \in \mathbb{R}^{n \times |\mathcal{S}|}$ is of full rank where $n < |\mathcal{S}|$; **(2)** For any ω, we have $\Phi\omega \neq \mathbf{1}$.*

In addition, we define a square matrix $A_\theta := \mathbb{E}_{s \sim d^{\pi_\theta}(\cdot), s' \sim \mathbb{P}^{\pi_\theta}(\cdot \mid s)}\{\phi(s)[\phi(s) - \phi(s')]^\top\}$. With Assumptions 2.1 and 3.3, $\omega^\top A_\theta \omega > 0$ holds for any ω, as proved in Tsitsiklis and Van Roy [1999]. Thus, this implies: **(1)** the matrix A_θ is invertible; **(2)** we have $\lambda_{\min}(A_\theta + A_\theta^\top) > 0$ (where $\lambda_{\min}$ denotes the smallest eigenvalue) due to $\omega^\top A_\theta \omega + \omega^\top A_\theta^\top \omega > 0$. We further assume that there exists a lower bound $\overline{\lambda} > 0$, such that $\lambda_{\min}(A_\theta + {A_\theta}^\top) \geq \overline{\lambda}$ holds for all θ. Then, we have the following proposition showing the existence and uniqueness of the solution to Equation (6). The detailed proof of this proposition can be found in Tsitsiklis and Van Roy [1999].

Algorithm 2 TD(0) Algorithm for Policy Evaluation

Input: Current policy π_θ.
Initialize: $\mu_0 = 0, \omega_0 = \mathbf{0}$.
for $\ell = 0, \ldots, L-1$ **do**
  Sample state s_ℓ, action a_ℓ, next state s'_ℓ, and $r_{\ell+1} = r(s_\ell, a_\ell, s'_\ell)$.
  Update μ by $\mu_{\ell+1} = \mu_\ell + \beta_{\mu,l} \cdot (r_{\ell+1} - \mu_\ell)$.
end for
Let $\overline{\mu}_L = \sum_{\ell=1}^{L} \mu_\ell / L$.
for $k = 0, \ldots, K-1$ **do**
  Sample state s_k, action a_k, next state s'_k, and $r_{k+1} = r(s_k, a_k, s'_k)$.
  Compute $\phi'_k := \phi(s'_k)$ and $\phi_k := \phi(s_k)$.
  Compute TD error: $\delta_k = r_{k+1} - \overline{\mu}_L + {\phi'_k}^\top \omega_k - \phi_k^\top \omega_k$.
  Update ω by $\omega_{k+1} = \Pi_\Omega(\omega_k + \beta_{\omega,k} \cdot \delta_k \cdot \phi_k)$.
end for
Let $\overline{\omega}_K = \sum_{k=1}^{K} \omega_k / K$.
Return: $\overline{\mu}_L$, $\overline{\omega}_K$.

Proposition 3.4 (Existence and Uniqueness of Solution)**.** *For the optimization problem in Equation* (6)*, under the Assumption 2.1 and Assumption 3.3, there exists a unique solution* $\{\omega_t^*, \mu_t^*\}$*, where* $\mu_t^* = J(\pi_{\theta_t}) = \mathbb{E}_{(s,a,s')\sim\Upsilon_{\theta_t}}(r)$ *and* $\omega_t^* = A_{\theta_t}^{-1} \cdot \mathbb{E}_{(s,a,s')\sim\Upsilon_{\theta_t}}[\phi(r - \mu_t^*)]$.

In Algorithm 2, the projection operator Π_Ω can restrict each iterate of ω in a convex set Ω. The projection exists only for rigorous theoretical analysis, which is common in the analysis of stochastic approximation. In practice, this operation may not be necessary. We define the convex set Ω in the following proposition. Detailed proof is shown in Supplementary Material §C.

Proposition 3.5 (Radius of Projection Region)**.** *In Algorithm 2, the set* Ω *is a ball with radius* $\mathcal{R}_{\text{proj}}$*, where the radius is* $\mathcal{R}_{\text{proj}} = 4r_{\max}/\overline{\lambda}$*, by which we can the feasible set as* $\Omega = \{\omega \mid \|\omega\|_2 \le \mathcal{R}_{\text{proj}}\}$.

4 Theoretical Results

One challenge in studying the online actor-critic algorithm is to analyze the gradient bias induced by the state mixing in the Markov chain. Here we make the following geometrically mixing assumption.

Assumption 4.1 (Geometrically Mixing)**.** *There are constants* C_{mix} *and* $\alpha \in (0,1)$ *such that* $\sup_{s\in\mathcal{S},\theta\in\Theta} d_{TV}(\mathbb{P}^{\pi_\theta}(s_t = \cdot \mid s_0 = s), d^{\pi_\theta}(\cdot)) \le C_{\text{mix}}\alpha^t$*, where* $d_{TV}(P,Q)$ *is total variation distance for two distributions* P *and* Q.

Geometrically mixing depicts the data sampling dependency in policy evaluation given a policy π_θ. Based on the notion of geometrically mixing, we can provide the convergence for TD(0).

Lemma 4.2 (Convergence of updating μ in TD(0))**.** *Let* $\beta_{\mu,l} = 1/(\ell+1)$ *for all* $\ell = 0, 1, \ldots, L-1$*. Then, under policy* π_θ *for any given* $\theta \in \Theta$*, supposing* $\mu^*(\theta)$ *is the solution to* (5)*, under the Assumptions 2.1, 3.3, and 4.1, the convergence of the* μ *updating step in Algorithm 2 is*

$$\mathbb{E}(\overline{\mu}_L - \mu^*(\theta))^2 \lesssim \frac{r_{\max}^2}{L} + \frac{r_{\max}^2 \log^2 L}{L}.$$

Please refer to the Supplementary Material §D for the detailed proof. As we can see, the geometrically mixing can only introduce a logarithm term $\log^2 L$ to the upper bound.

Theorem 4.3 (Convergence of TD(0) Algorithm)**.** *Assume that* $\mathbb{E}(\overline{\mu}_L - \mu^*(\theta))^2 \le \varepsilon_\mu$*. Let* $\beta_{\omega,k} = 1/(\ell+1)$ *for all* $\ell = 0, 1, \ldots, L-1$ *and* $\mu_0 = 0$*. Then, for any given* $\theta \in \Theta$*, supposing* $\omega^*(\theta)$ *is the solution to* (5)*, under the Assumptions 2.1, 3.3, and 4.1, the convergence of Algorithm 2 is*

$$\mathbb{E}\|V_{\overline{\omega}_K} - V_{\omega^*(\theta)}\|_D^2 \le \mathbb{E}\|\overline{\omega}_K - \omega^*(\theta)\|_2^2 \lesssim \kappa^2 r_{\max}^2 \frac{1}{K} + \kappa^4 r_{\max}^2 \frac{\log^2 K}{K} + \kappa^2 r_{\max}\sqrt{\varepsilon_\mu},$$

where we define $\kappa := \max\{2/\overline{\lambda}, 1\}$.

Please refer to the Supplementary Material §E for the detailed proof. We can interpret the convergence of TD(0) as follows:

- **Error terms:** The upper bound is composed of the learning errors of μ, ω. Particularly, if K is sufficiently large, then the upper bound in Theorem 4.3 is dominated by $\sqrt{\varepsilon_\mu}$.
- **Influence of geometrically mixing:**The geometrically mixing of the Markov chain only introduces $\log^2(K)$ term in the upper bound, which does not hurt the convergence too much.
- **Convergence rate:** The convergence of V_ω and ω is $\widetilde{O}^1(K^{-1}) + \sqrt{\varepsilon_\mu}$. If we let the iteration number be correlated with K, namely $L = O(K)$, the convergence of V_ω and ω is dominated by $\widetilde{O}(K^{-1/2})$. If we let $L = O(K^2)$, the convergence of V_ω and ω is $\widetilde{O}(K^{-1})$.

Corollary 4.4. *At the t-th step of Algorithm 1, the estimation error of μ and ω can be interpreted as*

$$\mathbb{E}(\widehat{\mu}_t - \mu_t^*)^2 \lesssim \varepsilon_\mu = \frac{r_{\max}^2}{L} + \frac{r_{\max}^2 \log^2 L}{L},$$

$$\mathbb{E}\|V_{\widehat{\omega}_t} - V_{\omega_t^*}\|_D^2 \leq \mathbb{E}\|\widehat{\omega}_t - \omega_t^*\|_2^2 \lesssim \kappa^2 r_{\max}^2 \frac{1}{K} + \kappa^4 r_{\max}^2 \frac{\log^2 K}{K} + \kappa^2 r_{\max}\sqrt{\varepsilon_\mu}.$$

This corollary can be obtained by setting the values of $(\widehat{\mu}_t, \widehat{\omega}_t)$ to be the output of Algorithm 2.

Next, we will show the convergence of the overall actor-critic algorithm. Before presenting the convergence of Algorithm 1, we make several proper assumptions that will be used in our analysis and proof for the following Theorem 4.6.

Assumption 4.5. *To show the convergence rate, we make the following assumptions: **(1)** There exists an optimal value $J^* := J(\pi^*)$ such that $\pi^* = \operatorname{argmax}_\pi J(\pi)$ and $J^* \leq +\infty$;**(2)** Bounded score function: $\|\nabla \log \pi_\theta(a \mid s)\|_2 \leq G_\pi, \forall \theta, \forall s, \forall a$; **(3)** ρ-Lipschitz gradient of $J(\pi_\theta)$: $\|\nabla_\theta J(\pi_{\theta_1}) - \nabla_\theta J(\pi_{\theta_2})\|_2 \leq \rho\|\theta_1 - \theta_2\|_2$ for $\forall \theta_1, \forall \theta_2$; **(4)** Uniformly bounded variance: $\mathbb{E}_{(s,a,s')\sim\Upsilon_\theta}\|g_\theta(\theta; s, a, s') - \nabla_\theta J(\pi_\theta)\|_2^2 \leq \sigma^2, \forall \theta$.*

Based on the above assumptions, we eventually have the following convergence rate for Algorithm 1.

Theorem 4.6 (Convergence of Actor-Critic Algorithm)**.** *Suppose that estimation errors of the critic step in Algorithm 1 satisfies $\mathbb{E}(\widehat{\mu}_t - \mu_t^*)^2 \leq \varepsilon_\mu$ and $\mathbb{E}\|\widehat{\omega}_t - \omega_t^*\|_2^2 \leq \varepsilon_\omega$. Under the Assumptions 2.1, 3.3, 4.1 and 4.5, letting $\beta_{\theta,t} = 1/(6\rho\sqrt{t+1})$, the convergence rate of Algorithm 1 is*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla J(\pi_{\theta_t})\|_2^2 \lesssim \frac{\rho[J(\pi^*) - J(\pi_{\theta_0})] + \sigma^2}{\sqrt{T}} + G_\pi^2 \varepsilon_\mu + G_\pi^2 \varepsilon_\omega + G_\pi^2 \mathcal{E}_{\text{approx}}.$$

Please refer to the Supplementary Material §F for the detailed proof.

In this paper, we assume that $J(\pi_\theta)$ is in a general nonconvex form w.r.t. the parameter θ. Therefore, under some additional proper assumptions in Assumption 4.5, the goal of Theorem 4.6 is to show how the iterates $\{\theta_t\}_{t\geq 0}$ converges to some stationary point of $J(\pi_\theta)$ with some irremovable error. By Theorem 4.6, we have the following results:

- **Error terms:** The upper bound is basically composed of the learning errors of μ, ω, and θ, and the value function approximation error $\mathcal{E}_{\text{approx}}$. This approximation error is irremovable due to the inherent nonlinearity of the value function.
- **Convergence rate:** If the learning error of TD(0) algorithm in Algorithm 2 is ε_ω and ε_μ respectively for μ and ω, the convergence of Algorithm 1 is $O(T^{-1/2}) + \varepsilon_\omega + \varepsilon_\mu + \mathcal{E}_{\text{approx}}$. If we let iteration numbers in Algorithm 2 be correlated with T in Algorithm 1 such that $L = K = O(T)$, the convergence of Algorithm 1 is $\widetilde{O}(T^{-1/2}) + \mathcal{E}_{\text{approx}}$ with $\varepsilon_\omega = \widetilde{O}(T^{-1/2})$ and $\varepsilon_\mu = \widetilde{O}(T^{-1})$.
- **Domination of $\mathcal{E}_{\text{approx}}$:** If T is sufficiently large and $\varepsilon_\mu, \varepsilon_\omega$ are sufficiently small, the policy gradient will converge up to the irremovable error of value function approximation $\mathcal{E}_{\text{approx}}$, namely $1/T\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\pi_{\theta_t})\|_2^2 \lesssim \mathcal{E}_{\text{approx}}$ with $T \gtrsim \mathcal{E}_{\text{approx}}^{-2}$, $\varepsilon_\mu \lesssim \mathcal{E}_{\text{approx}}$, and $\varepsilon_\omega \lesssim \mathcal{E}_{\text{approx}}$.
- **Special case $\mathcal{E}_{\text{approx}} = 0$:** Consider special case that the state-value function is intrinsically in a linear form with the basis function $\phi(s)$. Under this setting, the approximation error $\mathcal{E}_{\text{approx}}$ will be 0. With similar discussion on the convergence rate above, we can show that the actor-critic algorithm will converge to the stationary point in the sense of expectation, namely $1/T\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\pi_{\theta_t})\|_2^2 \lesssim \widetilde{O}(T^{-1/2})$ if letting $L = K = O(T)$ such that $\varepsilon_\omega = \widetilde{O}(T^{-1/2})$ and $\varepsilon_\mu = \widetilde{O}(T^{-1})$.

[1]We use $\widetilde{O}$ to hide the dependency on logarithm terms.

References

J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

S. Bhatnagar. An actor–critic algorithm with function approximation for discounted cost constrained Markov Decision Processes. *Systems & Control Letters*, 59(12):760–766, 2010.

S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S. Sutton. Incremental natural actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 105–112, 2008.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

V. S. Borkar. *Stochastic approximation: A dynamical systems viewpoint*. Cambridge University Press, 2008.

V. S. Borkar and S. P. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

J. A. Boyan. Least-squares temporal difference learning. In *International Conference on Machine Learning*, pages 49–56. Morgan Kaufmann Publishers Inc., 1999.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

D. D. Castro and R. Meir. A convergent online single-time-scale actor-critic algorithm. *Journal of Machine Learning Research*, 11(Jan):367–410, 2010.

T. Chen, K. Zhang, G. B. Giannakis, and T. Başar. Communication-efficient distributed reinforcement learning. *arXiv preprint arXiv:1812.03239*, 2018.

B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. Finite sample analyses for td (0) with function approximation. In *AAAI Conference on Artificial Intelligence*, 2018.

C. Dann, G. Neumann, J. Peters, et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1049–1058. JMLR. org, 2017.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

P. Hansen, B. Jaumard, and G. Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, New York, NY, 2003.

S. Lange, T. Gabel, and M. Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 504–513. AUAI Press, 2015.

H. R. Maei. Gradient temporal-difference learning algorithms. 2011.

K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7):1180–1190, 2008.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pages 387–395, 2014.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

A. Sinha, P. Malo, and K. Deb. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and td learning. *arXiv preprint arXiv:1902.00923*, 2019.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44, 1988.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

A. Touati, P.-L. Bacon, D. Precup, and P. Vincent. Convergent tree backup and retrace with function approximation. In *International Conference on Machine Learning*, pages 4962–4971, 2018.

J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11): 1799–1808, 1999.

O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, Y. Wu, D. Yogatama, J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, and D. Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`, 2019.

Y. Wang, W. Chen, Y. Liu, Z.-M. Ma, and T.-Y. Liu. Finite sample analysis of the gtd policy evaluation algorithms in markov setting. In *Advances in Neural Information Processing Systems*, pages 5504–5513, 2017.

Z. Yang, K. Zhang, M. Hong, and T. Başar. A finite sample analysis of the actor-critic algorithm. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2759–2764. IEEE, 2018.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*, 2018.

S. Zou, T. Xu, and Y. Liang. Finite-sample analysis for sarsa and q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234*, 2019.