Provably Efficient Convergence of Primal-Dual Actor-Critic with Nonlinear Function Approximation

Extended Abstract

Jing Dong
The Chinese University of Hong Kong, Shenzhen
jingdong@link.cuhk.edu.cn

Li Shen
JD Explorer Academy
mathshenli@gmail.com

Yinggan Xu
The Chinese University of Hong Kong, Shenzhen
yingganxu@link.cuhk.edu.cn

Baoxiang Wang
The Chinese University of Hong Kong, Shenzhen
bxiangwang@cuhk.edu.cn

ABSTRACT

We study the convergence of the actor-critic algorithm with nonlinear function approximation under a nonconvex-nonconcave primal-dual formulation. Stochastic gradient descent ascent is applied with an adaptive proximal term for robust learning rates. We show the first efficient convergence result with primal-dual actor-critic with a convergence rate of $O\left(\sqrt{\frac{\ln(NdG^2)}{N}}\right)$ under Markovian sampling, where G is the element-wise maximum of the gradient, N is the number of iterations, and d is the dimension of the gradient. Our result is presented with only the Polyak-Łojasiewicz (PL) condition for the dual variable, which is easy to verify and applicable to a wide range of RL scenarios.

KEYWORDS

Reinforcement learning, convergence analysis

ACM Reference Format:

Jing Dong, Li Shen, Yinggan Xu, and Baoxiang Wang. 2023. Provably Efficient Convergence of Primal-Dual Actor-Critic with Nonlinear Function Approximation: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Actor-critic [1, 2, 13] is one of the most successful algorithms in reinforcement learning. The algorithm features an actor, which learns the optimal policy that maximizes the long-term expected reward through sequential interactions with the environment, and a critic, which learns to approximate a value function that evaluates the performance of a policy. Armed with recent developments in deep learning, the actor-critic algorithm gains empirical success in a variety of real applications [9, 10, 12]. However, the underlying theory and limits have yet to be fully understood. Most previous analyses have their limitations. Castro and Meir [4], Maei [16] establish asymptotic convergence in the original setting with an unknown sample complexity. Follow-up works that investigate finite-sample performance are conducted with two-timescale updates [7, 11, 19] or linear function approximation [21, 22], where the best-known convergence rate is established to be $O(\epsilon^{-2/3})$. It is left open to theoretically justify the actor-critic method's practical achievements in theory in its general setting.

We study a *single-timescale* variant of the actor-critic algorithm with *nonlinear function approximation* based on a minimax optimization formulation that combines the objectives for actor and critic [5]. With our formulation, a convergence rate of $O\left(\sqrt{\frac{\ln(NdG^2)}{N}}\right)$ under Markovian sampling and adaptive gradient is shown, where N is the number of total iterations, d is the dimension of the gradient, and G is the element-wise maximum value of the gradient. This implies a sample complexity of $\tilde{O}(\epsilon^{-2})$ with a constant batch size that is independent of N and ϵ. Our theorems are under Polyak-Łojasiewicz (PL) condition with respect to the dual variable, which is a much weaker assumption than the Minty Variational Inequality (MVI) commonly seen in the nonconvex-nonconcave optimization literature [14, 15].

2 PRELIMINARIES

We consider a discounted Markov decision process (MDP) denoted by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability kernel such that given a state-action pair (s, a), it returns a probability distribution $s' \sim \mathbb{P}(\cdot \mid s, a)$ of the next state, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and γ is the discount factor.

The goal of reinforcement learning is to learn a policy π, which takes $s \in \mathcal{S}$ as an input and outputs a distribution $a \sim \pi(\cdot \mid s)$ over the action space $\mathcal{A}$, to maximize the expected cumulative discounted reward $\mathbb{E}_{s_0}\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$, where $s_0 \sim \mu_0$ is a given initial state distribution.

To evaluate the performance of the policy, the value function is defined to measure the long-term expected cumulative discounted reward as $V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s\right]$. Let V^* be the optimal value function such that $V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s\right]$. The Bellman optimality equation states that

$$V^*(s_t) = \Gamma V^*(s_t) = \max_{a \in \mathcal{A}} \left\{R(s_t, a_t) + \gamma^t \mathbb{E}_{s_{t+1}}\left[V^*(s_{t+1})\right]\right\} . \quad (1)$$

Equation (1) can then be formulated into the following linear program (LP) [3],

$$\begin{aligned} V^* = \underset{V}{\text{minimize}} \quad & (1 - \gamma^t)\mathbb{E}_{s_t}\left[V(s_t)\right] \\ \text{subject to} \quad & V(s_t) \geq R(s_t, a_t) + \gamma^t \mathbb{E}_{s_{t+1}}\left[V(s_{t+1})\right] . \end{aligned} \quad (2)$$

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Without loss of generality, we assume that the linear program is feasible, i.e., there exists an optimal policy for the given MDP.

When the strong duality holds, by [5], the equivalent saddle point problem (3) can be jointly optimized to learn both the policy and value functions. This approach of learning approximate policy and value functions simultaneously is known as the actor-critic method. The formulation is

$$\min_{V}\max_{\alpha,\pi} L_k(V,\alpha,\pi) = \min_{V}\max_{\alpha,\pi} (1-\gamma^{k+1})\mathbb{E}_\mu[V(s_t)] + \sum_{(s_t,a_t)_{t=0}^{k}, s_{k+1}} \alpha(s_t)\zeta(s_t,a_t,s_{t+1}), \tag{3}$$

where $\alpha : \mathcal{S} \to \Delta(\mathcal{S}), \pi : \mathcal{S} \to \Delta(\mathcal{A})$, $\zeta(s_t,a_t,s_{t+1}) = \prod_{t=0}^{k} \pi(a_t \mid s_t)P(s_{t+1} \mid s_t,a_t)\delta((s_t,a_t)_{t=0}^{k}, s_{k+1})$, and $\delta((s_t,a_t)_{t=0}^{k}, s_{k+1}) = \sum_{t=0}^{k} \gamma^t R(s_t,a_t) + \gamma^{k+1}V(s_{k+1}) - V(s_t)$.

Assume that α, π, V are parameterized by u, θ, ω, respectively. Let $\nabla_u L_k, \nabla_\theta L_k, \nabla_\omega L$ denote the gradients of Equation (3) with respect to each parameter. We design a variant of SGDA with adaptive gradients, which dynamically incorporates the history of the gradients to construct more informative updates. The algorithm is described in Algorithm 1

Algorithm 1 Adaptive SGDA (ASGDA)

```
Input: Learning rates $\eta_\omega, \eta_z = (\eta_u, \eta_\theta)$, batch size $M$, $H_0 = I$, $z = (u,\theta)$, $\hat{G}_z = G + \xi\left(D + \frac{2D}{1-\gamma}\right)^2 \cdot (D_u^2 + D_\theta^2)$, $\hat{G}_\theta = G + \xi(D + 2D_\omega)^2$
for $k = 1, \ldots, N$ do
    Start from $s \sim \alpha_k(s)$ where $\alpha_k$ is parametrized by $u_k$, collect samples $\tau_k = \{s_t, a_t, r_t, s_{t+1}\}_{t=0}^{M}$ following policy $\pi_k$ parametrized by $\hat{\theta}_{k-1}$
    $\hat{g}_\omega(\hat{\omega}_k, \hat{z}_k) = \nabla_\omega L(\hat{\omega}_k, \hat{u}_k, \hat{\theta}_k, \tau_k)$,   $\hat{g}_z(\hat{\omega}_k, \hat{z}_k) = \nabla_z L(\hat{\omega}_k, \hat{u}_k, \hat{\theta}_k, \tau_k)$
    $\hat{\omega}_k = \omega_{k-1} - \eta_\omega \left(I + \sqrt{\hat{H}_{\omega,k-1}^{-1}}\right) \hat{g}_\omega(\hat{\omega}_{k-1}, \hat{z}_{k-1})$
    $\omega_k = \omega_{k-1} - \eta_\omega \left(I + \sqrt{\hat{H}_{\omega,k}^{-1}}\right) \hat{g}_\omega(\hat{\omega}_k, \hat{z}_k)$
    $\hat{z}_k = z_{k-1} + \eta_z \left(I + \sqrt{\hat{H}_{z,k-1}^{-1}}\right) \hat{g}_z(\hat{\omega}_{k-1}, \hat{z}_{k-1})$
    $z_k = z_{k-1} + \eta_z \left(I + \sqrt{\hat{H}_{z,k}^{-1}}\right) \hat{g}_z(\hat{\omega}_k, \hat{z}_k)$
    $\hat{g}_{\omega,0:k} = \frac{1}{\sqrt{2}\hat{G}_\omega}[\hat{g}_{\omega,0:k-1}, \hat{g}_\omega(\hat{\omega}_k, \hat{z}_k)]$
    $\hat{h}_{\omega,k,i} = \|g_{\omega,0:k,i}\|^2, i = 1, \ldots, d$, $\hat{H}_{\omega,k} = \mathbf{Diag}(\hat{h}_{\omega,k}) + \frac{1}{2}I$
    $\hat{g}_{z,0:k} = \frac{1}{\sqrt{2}\hat{G}_z}[\hat{g}_{z,0:k-1}, \hat{g}_z(\hat{\omega}_k, \hat{z}_k)]$
    $\hat{h}_{z,k,i} = \|\hat{g}_{z,k,i}\|^2, i = 1, \ldots, d$, $\hat{H}_{z,k} = \mathbf{Diag}(\hat{h}_{z,k-1}) + \frac{1}{2}I$
end for
```

3 CONVERGENCE ANALYSIS

Before we present the main theorem, we first discuss the assumptions necessary for the results. Most of the previous analyses on nonconvex-nonconcave optimization problems utilize the MVI inequality assumption [6, 14, 15], which is unrealistic in real applications. Instead, we consider one-sided Polyak-Łojasiewicz (PL) inequality for the dual variables only, which is relatively weaker compared to MVI.

Assumption 3.1 (PL condition for dual variables). *$L(\omega, z)$ is assumed to satisfy Polyak-Łojasiewicz (PL) condition with respect to z such that $\forall\omega \in \mathbb{R}^d$ and for some constant $\mu, \frac{1}{2}\|\nabla_z L(\omega,z)\|^2 \geq \mu(L(\omega, f^*(z)) - L(\omega,z))$, holds for all $z = (u,\theta) \in \mathbb{R}^d \times \mathbb{R}^d$.*

Beyond the PL condition, we also assume that the gradients are Lipschitz continuous and bounded. These assumptions are common among the optimization literature [8, 17].

Assumption 3.2 (Lipschitz continuity and boundedness of the gradient). *There exists a constant C such that for all $(\omega,z), (\omega',z')$ $\|\nabla L(\omega,z) - \nabla L(\omega',z')\| \leq C\|(\omega,z) - (\omega',z')\|$, where $\nabla L(\omega,z) = (\nabla_\omega L(\omega,z), \nabla_z L(\omega,z))$. There also exist constants D_u, D_θ, D_ω such that for all (u,θ,ω), $\|\nabla_u \log(\alpha_u(s))\| \leq D_u$, $\|\nabla_\theta \log(\pi_\theta(a|s))\| \leq D_\theta$, $\|\nabla_\omega V_\omega(s)\| \leq D_\omega$.*

We also need the following assumption regarding the underlying MDP, which is common for analyses under Markovian sampling and can be satisfied by time-homogeneous Markov chains with finite state space [18, 20].

Assumption 3.3 (Geometric convergence rate of MDP). *The MDP is irreducible and aperiodic for all π, and there exist constants $\xi > 0$ and $\rho \in (0,1)$ such that for all π and $t \geq 0$, $\sup_{s\in\mathcal{S}} \|P(s_t,\cdot) - \kappa(\cdot)\| \leq \xi\rho^t$, where $\kappa(\cdot)$ is the stationary distribution of the Markov chain induced by policy π.*

We further assume the rewards are bounded, to prevent the value function to go unbounded.

Assumption 3.4 (Bounded reward). *There exists a constant D such that $|R(s,a)| \leq D, \forall s \in \mathcal{S}, a \in \mathcal{A}$.*

To measure the convergence of our algorithm, we consider the first-order stationary point of the envelope function $\Phi(\omega)$, $\Phi(\omega) = L(\omega, f^*(\omega))$, where $f^*(\omega) = \operatorname{argmax}_z L(\omega,z)$. As the strong duality holds for our formulation and ω parametrized the value function, the envelop function indirectly evaluates the convergence to the local optimal value function.

Armed with the above assumptions, we give the following convergence guarantee for Algorithm 1.

Theorem 3.1. *Under appriopriate choices of η_z, η_ω and Assumption 3.1, 3.2, 3.3, 3.4, we have $\sum_{k=2}^{N} \mathbb{E}\left[\|\nabla_\omega \Phi(\omega_k)\|^2\right]$ $= O\left(\frac{\max\{\hat{G}_\omega^2, \hat{G}_z^2\} \ln\left(dNG^2 + \frac{1}{M}\sum_{m=1}^{M} \xi\rho^{2m}\right)}{N}\right)$, where N is the number of iterations and $\hat{G}_\omega$, $\hat{G}_z$ are constants.*

4 CONCLUSION AND FUTURE WORKS

We investigate the primal-dual formulation of the actor-critic method in reinforcement learning. We presented the first finite-sample analysis for single-scale algorithms and nonlinear function approximation. Under Markovian sampling and adaptive gradients, we establish a convergence rate of $\tilde{O}(\epsilon^{-2})$. This guarantee is under PL conditions for only the dual variables.

ACKNOWLEDGMENTS

This work is supported by the Major Science and Technology Innovation 2030 “Brain Science and Brain-like Research” key project (No. 2021ZD0201405). Jing Dong and Baoxiang Wang are partially supported by the National Natural Science Foundation of China (62106213, 72150002) and Shenzhen Science and Technology Program (RCBS20210609104356063, JCYJ20210324120011032).

REFERENCES

[1] Andrew Gehret Barto, Richard S Sutton, and Charles W Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* (1983), 834–846.

[2] Andrew Gehret Barto, Richard S Sutton, and Christopher JCH Watkins. 1989. Learning and sequential decision making. *COINS Technical Report 89-95* (1989).

[3] Dimitri Bertsekas. 2000. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont.

[4] Dotan Di Castro and Ron Meir. 2010. A convergent online single time scale actor critic algorithm. *The Journal of Machine Learning Research* 11 (2010), 367–410.

[5] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. 2018. Boosting the actor with dual critic. In *International Conference on Learning Representations*.

[6] Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. 2021. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*.

[7] Thinh T Doan. 2021. Finite-time convergence rates of nonlinear two-time-scale stochastic approximation under Markovian noise. *arXiv preprint arXiv:2104.01627* (2021).

[8] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, 7 (2011).

[9] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*.

[10] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*.

[11] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. 2020. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170* (2020).

[12] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*.

[13] Vijay R Konda and John N Tsitsiklis. 1999. Actor-citic agorithms. In *International Conference on Neural Information Processing Systems*.

[14] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. 2018. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207* (2018).

[15] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. 2019. Towards Better Understanding of Adaptive Gradient Algorithms in Generative Adversarial Nets. In *International Conference on Learning Representations*.

[16] Hamid Reza Maei. 2018. Convergent actor-critic algorithms under off-policy training and function approximation. *arXiv preprint arXiv:1802.07842* (2018).

[17] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. 2017. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*.

[18] Tao Sun, Han Shen, Tianyi Chen, and Dongsheng Li. 2020. Adaptive temporal difference learning with linear function approximation. *arXiv preprint arXiv:2002.08537* (2020).

[19] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. 2020. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350* (2020).

[20] Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. 2020. Non-asymptotic convergence of Adam-type reinforcement learning algorithms under Markovian sampling. *arXiv preprint arXiv:2002.06286* (2020).

[21] Tengyu Xu, Zhe Wang, and Yingbin Liang. 2020. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557* (2020).

[22] Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. 2021. Doubly Robust Off-Policy Actor-Critic: Convergence and Optimality. In *International Conference on Machine Learning*.