Finite-Time Analysis of Single-Timescale Actor-Critic

Xuyang Chen, Lin Zhao

Department of Electrical & Computer Engineering
National University of Singapore

November 10, 2023

- **Motivation**: we study the finite-time convergence of single-timescale actor-critic algorithm under the Markovian sampling scheme with infinite state space and average reward setting.

- **Motivation**: we study the finite-time convergence of single-timescale actor-critic algorithm under the Markovian sampling scheme with infinite state space and average reward setting.
- **Challenge**: how to control the highly coupled error propagation between reward, critic, and actor in this setting?

Background

- **Motivation**: we study the finite-time convergence of single-timescale actor-critic algorithm under the Markovian sampling scheme with infinite state space and average reward setting.
- **Challenge**: how to control the highly coupled error propagation between reward, critic, and actor in this setting?
- **Idea**: keep track of these errors to establish an interconnected iteration system and solve them simultaneously.

We consider the standard Markov Decision Process (MDP) characterized by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. We consider a finite action space $|\mathcal{A}| < \infty$, whereas the state space can be either a finite set or an (unbounded) real vector space $\mathcal{S} \subset \mathbb{R}^n$. $\mathcal{P}(s_{t+1}|s_t, a_t) \in [0, 1]$ denotes the transition kernel. We consider a bounded reward $r : \mathcal{S} \times \mathcal{A} \to [-U_r, U_r]$, which is a function of the state s and action a. A policy $\pi_{\boldsymbol{\theta}}(\cdot|s) \in \mathbb{R}^{|\mathcal{A}|}$ parameterized by $\boldsymbol{\theta}$ is defined as a mapping from a given state to a probability distribution over actions.

The RL problem of consideration aims to find a policy $\pi_{\boldsymbol{\theta}}$ that maximizes the infinite-horizon time-average reward, which is given by

$$J(\boldsymbol{\theta}) := \lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\theta}} \frac{\sum_{t=0}^{T-1} r(s_t, a_t)}{T} = \mathop{\mathbb{E}}_{s \sim \mu_{\boldsymbol{\theta}}, a \sim \pi_{\boldsymbol{\theta}}} [r(s, a)],$$

where the expectation $\mathbb{E}_{\boldsymbol{\theta}}$ is over the Markov chain under the policy $\pi_{\boldsymbol{\theta}}$, and $\mu_{\boldsymbol{\theta}}$ denotes the stationary state distribution induced by $\pi_{\boldsymbol{\theta}}$.

Algorithm

We analyze the following algorithm for finding optimal policy $\pi_{\boldsymbol{\theta}}$.

Algorithm Single-timescale Actor-Critic

Input initial actor parameter $\boldsymbol{\theta}_0$, initial critic parameter $\boldsymbol{\omega}_0$, initial reward estimator η_0, stepsize α_t for actor, β_t for critic, and γ_t for reward estimator.
Draw s_0 from some initial distribution
for $t = 0, 1, 2, \cdots, T-1$ **do**
    Take action $a_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)$
    Observe next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ and reward $r_t = r(s_t, a_t)$
    $\delta_t = r_t - \eta_t + \boldsymbol{\phi}(s_{t+1})^\top \boldsymbol{\omega}_t - \boldsymbol{\phi}(s_t)^\top \boldsymbol{\omega}_t$
    $\eta_{t+1} = \eta_t + \gamma_t(r_t - \eta_t)$
    $\boldsymbol{\omega}_{t+1} = \Pi_{U_{\boldsymbol{\omega}}}(\boldsymbol{\omega}_t + \beta_t \delta_t \boldsymbol{\phi}(s_t))$
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \delta_t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)$
end for

Algorithm

We analyze the following algorithm for finding optimal policy $\pi_{\boldsymbol{\theta}}$.

Algorithm Single-timescale Actor-Critic

Input initial actor parameter $\boldsymbol{\theta}_0$, initial critic parameter $\boldsymbol{\omega}_0$, initial reward estimator η_0, stepsize α_t for actor, β_t for critic, and γ_t for reward estimator.
Draw s_0 from some initial distribution
for $t = 0, 1, 2, \cdots, T-1$ **do**
    Take action $a_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|s_t)$
    Observe next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ and reward $r_t = r(s_t, a_t)$
    $\delta_t = r_t - \eta_t + \boldsymbol{\phi}(s_{t+1})^\top \boldsymbol{\omega}_t - \boldsymbol{\phi}(s_t)^\top \boldsymbol{\omega}_t$
    $\eta_{t+1} = \eta_t + \gamma_t(r_t - \eta_t)$
    $\boldsymbol{\omega}_{t+1} = \Pi_{U_{\boldsymbol{\omega}}}(\boldsymbol{\omega}_t + \beta_t \delta_t \boldsymbol{\phi}(s_t))$
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \delta_t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)$
end for

- Note that the “single-timescale” refers to the fact that the stepsizes $\alpha_t, \beta_t, \gamma_t$ are only constantly proportional to each other.

Assumption 1 (Exploration)

$\boldsymbol{A}_{\boldsymbol{\theta}} := \mathbb{E}_{(s,a,s')}[\phi(s)(\phi(s') - \phi(s))^\top)]$ with $s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$ is negative definite and its maximum eigenvalue can be upper bounded by $-\lambda$.

Assumption 1 (Exploration)

$\boldsymbol{A}_{\boldsymbol{\theta}} := \mathbb{E}_{(s,a,s')}[\phi(s)(\phi(s') - \phi(s))^{\top})]$ with $s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$ is negative definite and its maximum eigenvalue can be upper bounded by $-\lambda$.

Assumption 2 (Uniform ergodicity)

For a Markov chain generated by $\pi_{\boldsymbol{\theta}}$ and $\mathcal{P}$, there exists $m > 0$ and $\rho \in (0,1)$ such that $d_{TV}(\mathbb{P}(s_\tau \in \cdot|s_0 = s), \mu_{\boldsymbol{\theta}}(\cdot)) \leq m\rho^\tau, \forall \tau \geq 0, \forall s \in \mathcal{S}$.

Assumption 1 (Exploration)

$\boldsymbol{A}_{\boldsymbol{\theta}} := \mathbb{E}_{(s,a,s')}[\phi(s)(\phi(s') - \phi(s))^\top)]$ with $s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$ is negative definite and its maximum eigenvalue can be upper bounded by $-\lambda$.

Assumption 2 (Uniform ergodicity)

For a Markov chain generated by $\pi_{\boldsymbol{\theta}}$ and $\mathcal{P}$, there exists $m > 0$ and $\rho \in (0,1)$ such that $d_{TV}(\mathbb{P}(s_\tau \in \cdot|s_0 = s), \mu_{\boldsymbol{\theta}}(\cdot)) \leq m\rho^\tau, \forall \tau \geq 0, \forall s \in \mathcal{S}$.

Assumption 3 (Lipschitz continuity of policy)

There exist constants B, L_l, L_π such that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, it holds that: i)$\|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \leq B$; ii)$\|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \leq L_l\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$; iii)$|\pi_{\boldsymbol{\theta}_1}(a|s) - \pi_{\boldsymbol{\theta}_2}(a|s)| \leq L_\pi\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.

Assumption 1 (Exploration)

$\boldsymbol{A}_{\boldsymbol{\theta}} := \mathbb{E}_{(s,a,s')}[\phi(s)(\phi(s') - \phi(s))^\top)]$ with $s \sim \mu_{\boldsymbol{\theta}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)$ is negative definite and its maximum eigenvalue can be upper bounded by $-\lambda$.

Assumption 2 (Uniform ergodicity)

For a Markov chain generated by $\pi_{\boldsymbol{\theta}}$ and $\mathcal{P}$, there exists $m > 0$ and $\rho \in (0,1)$ such that $d_{TV}(\mathbb{P}(s_\tau \in \cdot|s_0 = s), \mu_{\boldsymbol{\theta}}(\cdot)) \leq m\rho^\tau, \forall \tau \geq 0, \forall s \in \mathcal{S}$.

Assumption 3 (Lipschitz continuity of policy)

There exist constants B, L_l, L_π such that for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, it holds that: i)$\|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\| \leq B$; ii)$\|\nabla \log \pi_{\boldsymbol{\theta}_1}(a|s) - \nabla \log \pi_{\boldsymbol{\theta}_2}(a|s)\| \leq L_l\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$; iii)$|\pi_{\boldsymbol{\theta}_1}(a|s) - \pi_{\boldsymbol{\theta}_2}(a|s)| \leq L_\pi\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.

Assumption 4 (Lipschitz continuity of stationary distribution)

For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, there exists constant L_μ such that $\|\nabla \mu_{\boldsymbol{\theta}} - \nabla \mu_{\boldsymbol{\theta}'}\| \leq L_\mu\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, where $\mu_{\boldsymbol{\theta}}(s)$ is the stationary distribution under the policy $\pi_{\boldsymbol{\theta}}$.

Theorem 5 (Markovian sampling)

Consider Algorithm 1 with $\alpha_t = \frac{c}{\sqrt{T}}, \beta_t = \gamma_t = \frac{1}{\sqrt{T}}$, *where* c *is a constant depending on problem parameters. Suppose Assumptions 1-4 hold, we have for* $T \geq 2\tau_T$,

$$\frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}y_t^2 = \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + \mathcal{O}(\epsilon_{\mathrm{app}}),$$

$$\frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}\|\boldsymbol{z}_t\|^2 = \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + \mathcal{O}(\epsilon_{\mathrm{app}}),$$

$$\frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2 = \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + \mathcal{O}(\epsilon_{\mathrm{app}}).$$

- ϵ_{app} is the critic approximation error.
- $y_t := \eta_t - J(\boldsymbol{\theta}_t)$ and $\boldsymbol{z}_t := \boldsymbol{\omega}_t - \boldsymbol{\omega}^*(\boldsymbol{\theta}_t)$ measure the reward estimation error and critic error, respectively.
- $\tau_T = \frac{\log m\rho^{-1}}{\log \rho^{-1}} + \frac{\log T}{2 \log \rho^{-1}} = \mathcal{O}(\log T)$ represents the mixing time of an ergodic Markov chain.
- To obtain an ϵ-approximate stationary point, it takes a number of $\widetilde{\mathcal{O}}(\epsilon^{-2})$ samples for Markovian sampling and $\mathcal{O}(\epsilon^{-2})$ for i.i.d. sampling, which matches the state-of-the-art performance of SGD on non-convex optimization problems.

- Reward Estimation Error: from the reward estimator update rule in Line 7 of Algorithm 1, we decompose the reward estimation error into:

$$
\begin{aligned}
y_{t+1}^2 = &(1-2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - J(\boldsymbol{\theta}_t)) + 2y_t(J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1})) \\
&+ (J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1}) + \gamma_t(r_t - \eta_t))^2.
\end{aligned} \tag{1}
$$

- Reward Estimation Error: from the reward estimator update rule in Line 7 of Algorithm 1, we decompose the reward estimation error into:

$$
\begin{aligned}
y_{t+1}^2 = &(1-2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - J(\boldsymbol{\theta}_t)) + 2y_t(J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1})) \\
&+ (J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1}) + \gamma_t(r_t - \eta_t))^2.
\end{aligned} \tag{1}
$$

- Critic Error: from the critic update rule in Line 8 of Algorithm 1, we decompose the squared critic error into

$$
\begin{aligned}
\|\boldsymbol{z}_{t+1}\|^2 = &\|\boldsymbol{z}_t\|^2 + 2\beta_t\langle \boldsymbol{z}_t, \bar{g}(\boldsymbol{\omega}_t, \boldsymbol{\theta}_t)\rangle + 2\beta_t\Psi(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \\
&+ 2\beta_t\langle \boldsymbol{z}_t, \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)\rangle + 2\langle \boldsymbol{z}_t, \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\rangle \\
&+ \|\beta_t(g(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) + \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)) + \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\|^2.
\end{aligned} \tag{2}
$$

Proof Sketch

- Reward Estimation Error: from the reward estimator update rule in Line 7 of Algorithm 1, we decompose the reward estimation error into:

$$\begin{aligned} y_{t+1}^2 = (1-2\gamma_t)y_t^2 + 2\gamma_t y_t(r_t - J(\boldsymbol{\theta}_t)) + 2y_t(J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1})) \\ + (J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1}) + \gamma_t(r_t - \eta_t))^2. \end{aligned} \tag{1}$$

- Critic Error: from the critic update rule in Line 8 of Algorithm 1, we decompose the squared critic error into

$$\begin{aligned} \|\boldsymbol{z}_{t+1}\|^2 = &\|\boldsymbol{z}_t\|^2 + 2\beta_t\langle \boldsymbol{z}_t, \bar{g}(\boldsymbol{\omega}_t, \boldsymbol{\theta}_t)\rangle + 2\beta_t\Psi(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) \\ &+ 2\beta_t\langle \boldsymbol{z}_t, \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)\rangle + 2\langle \boldsymbol{z}_t, \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\rangle \\ &+ \|\beta_t(g(O_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t) + \Delta g(O_t, \eta_t, \boldsymbol{\theta}_t)) + \boldsymbol{\omega}_t^* - \boldsymbol{\omega}_{t+1}^*\|^2. \end{aligned} \tag{2}$$

- Policy Gradient Norm (Actor Error): from the actor update rule in Line 9 of Algorithm 1, we bound the policy gradient norm by

$$\begin{aligned} \|\nabla J(\boldsymbol{\theta}_t)\|^2 \le &\frac{1}{\alpha_t}(J(\boldsymbol{\theta}_{t+1}) - J(\boldsymbol{\theta}_t)) - \langle \nabla J(\boldsymbol{\theta}_t), \Delta h(O_t, \eta_t, \boldsymbol{\omega}_t, \boldsymbol{\theta}_t)\rangle \\ &- \langle \nabla J(\boldsymbol{\theta}_t), \mathbb{E}_{O_t'}[\Delta h'(O_t', \boldsymbol{\theta}_t)]\rangle \\ &+ \Theta(O_t, \boldsymbol{\theta}_t) + \frac{L_{J'}}{2}\alpha_t\|\delta_t\nabla \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t)\|^2. \end{aligned} \tag{3}$$

Taking expectation of and summing (1),(2),and (3) from τ_T to $T-1$, we define $Y_T = \frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}y_t^2$, $Z_T = \frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}\|\boldsymbol{z}_t\|^2$, $G_T = \frac{1}{T-\tau_T}\sum_{t=\tau_T}^{T-1}\mathbb{E}\|\nabla J(\boldsymbol{\theta}_t)\|^2$.
By analysing each error term in (1),(2), and (3), we obtain the following **interconnected iteration system:**

$$Y_T \le \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + l_1\sqrt{Y_T G_T},$$

$$Z_T \le \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + \mathcal{O}(\epsilon_{\mathrm{app}}) + l_2\sqrt{Y_T Z_T} + l_3\sqrt{Z_T(2Y_T + 8Z_T)},$$

$$G_T \le \mathcal{O}(\frac{\log^2 T}{\sqrt{T}}) + \mathcal{O}(\epsilon_{\mathrm{app}}) + l_4\sqrt{G_T(2Y_T + 8Z_T)},$$

where l_1, l_2, l_3, l_4 are positive constants. By solving the above system of inequalities, we further prove that if $l_1(1 + 2l_4^2 + 8l_4^2(2l_2^2 + l_3)) \le 1$ and $16l_3 \le 1$, which can be easily satisfied by choosing the following stepsize ratio $c = \min\{\frac{\lambda}{32BL_*}, \frac{\lambda^2}{G(\lambda^2+3B^2\lambda^2+64B^2)}\}$, then Y_T, Z_T, G_T converge at a rate of $\mathcal{O}(\frac{\log^2 T}{\sqrt{T}})$. Therefore, we conclude our proof.

Contribution

Table: Comparison with related single-timescale actor-critic algorithms

Reference	Setting		Sampling		Sample Complexity
	State Space	Reward	Actor	Critic	
Olshevsky & Gharesifard	Finite	Discounted	i.i.d.	i.i.d.	$\mathcal{O}(\epsilon^{-2})$
Chen et al. (2021)	Infinite	Discounted	i.i.d.	i.i.d.	$\mathcal{O}(\epsilon^{-2})$
This Paper	Infinite	Average	Markovian	Markovian	$\tilde{\mathcal{O}}(\epsilon^{-2})$

- We for the first time show the finite-time analysis of single-timescale actor-critic under the Markovian sampling setting.
- We develop a new analysis framework that can be potentially applied to analyze other single-timescale stochastic approximation algorithms.

Thank You !