Two-Timescale Critic-Actor for Average Reward MDPs with Function Approximation

Prashansa Panda and Shalabh Bhatnagar

Indian Institute of Science, Bengaluru, India.
prashansap@iisc.ac.in; shalabh@iisc.ac.in

Abstract

Several recent works have focused on carrying out non-asymptotic convergence analyses for AC algorithms. Recently, a two-timescale critic-actor algorithm has been presented for the discounted cost setting in the look-up table case where the timescales of the actor and the critic are reversed and only asymptotic convergence shown. In our work, we present the first two-timescale critic-actor algorithm with function approximation in the long-run average reward setting and present the first finite-time non-asymptotic as well as asymptotic convergence analysis for such a scheme. We obtain optimal learning rates and prove that our algorithm achieves a sample complexity that can be made arbitrarily close to that of single-timescale AC and clearly better than the one obtained for two-timescale AC in a similar setting. A notable feature of our analysis is that we present the asymptotic convergence analysis of our scheme in addition to the finite-time bounds that we obtain and show the almost sure asymptotic convergence of the (slower) critic recursion to the attractor of an associated differential inclusion with actor parameters corresponding to local maxima of a perturbed average reward objective. We also show the results of numerical experiments on three benchmark settings and observe that our critic-actor algorithm performs the best amongst all algorithms.

1 Introduction

Actor-Critic (AC) methods have proved to be efficient in solving many reinforcement learning (RL) tasks. Actor-only methods such as REINFORCE suffer from high variance during the estimation of the policy gradient whereas critic-only methods like Q-learning are efficient in the tabular setting but can diverge when function approximation is used. AC methods try to circumvent these problems by combining both policy- and value-based methods to solve RL problems. In these approaches, the goal of the actor is to learn the optimal policy using the value updates provided by the critic, while the goal of the critic is to learn the value function for a policy prescribed by the actor. One obtains stable behavior of such algorithms through a difference in timescales that we explain in more detail below.

The AC framework is designed to mimic the policy iteration (PI) procedure for Markov decision processes (Puterman 2014). The AC algorithms incorporate two-timescale coupled stochastic recursions with the learning rate of the actor typically converging to zero at a rate faster than that of the critic. The timescale separation in two-timescale stochastic approximation algorithms such as AC is critical in ensuring stability of the recursions and their almost sure convergence. This is because from the viewpoint of the faster timescale, the slower recursion appears to be quasi-static while from the viewpoint of the slower timescale, the faster recursion appears to have converged. This helps the AC scheme to emulate policy iteration and thereby converge to the optimal policy. Asymptotic convergence analyses of two-timescale AC schemes are largely available via the ordinary differential equation (ODE) based approach. In (Bhatnagar, Borkar, and Guin 2023), the critic-actor (CA) algorithm was proposed, in the lookup table setting for the infinite-horizon discounted cost criterion, where the roles of the actor and the critic were reversed by swapping their timescales. The resulting procedure is seen to track value iteration instead of policy iteration.

In this paper, we carry this idea forward and present, for the first time, a critic-actor algorithm with function approximation and for the long-run average (and not discounted) reward setting. We then carry out detailed asymptotic and non-asymptotic convergence analyses of the same. Our algorithm runs temporal difference learning on the slower timescale to estimate the critic updates and stochastic policy gradient on the faster timescale for the actor. We prove that this algorithm emulates an approximate value iteration scheme. Our paper plugs in an important gap that existed previously by studying a new class of algorithms obtained by merely reversing the timescales of the actor and the critic. Our finite-time analysis shows that the two-timescale CA algorithm has a better sample complexity when compared with the two-timescale AC algorithm.

In our algorithm, even though there are three recursions, the average reward and actor recursions together proceed on the same timescale that is faster than the timescale of the critic update. Notice the difference of our scheme with AC algorithms for average-reward MDPs such as those in (Wu et al. 2022; Bhatnagar et al. 2009), where the average-reward recursion proceeds on the same (faster) timescale as the critic while the actor recursion proceeds slower. We use linear function approximation for the critic recursion and a policy gradient actor. We perform the non-asymptotic analysis of

this algorithm and obtain its sample complexity. In addition, we prove that the scheme remains asymptotically stable and is almost surely convergent to the attractors of an underlying differential inclusion. Our analysis helps us in getting the optimised learning rates for the actor and the critic recursions. Finally, we show numerical performance comparisons of our algorithm with the AC and a few other algorithms over three different OpenAI Gym environments and observe that the CA algorithm shows the best performance amongst all algorithms considered, though by small margins. In terms of the training time performance, CA is better than all algorithms except DQN on all three environments, and in fact, it takes about half the training time on two of the environments.

Main Contributions: (*a*) We present the first critic-actor (CA) algorithm with linear function approximation for the long-run average-reward criterion where the critic runs on a slower timescale than the actor.
(*b*) We carry out the first finite-time analysis of the two-timescale CA algorithm wherein we present finite-time bounds for the critic error, actor error and the average reward estimation error, respectively. In particular, we obtain a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-(2+\delta)})$ with $\delta > 0$ arbitrarily close to zero, for the mean squared error of the critic to be upper bounded by ϵ. This is better than the sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ obtained by the two-timescale AC algorithm of (Wu et al. 2022) and can be brought as close as possible to the sample complexity of the recently studied single-timescale AC schemes (Olshevsky and Gharesifard 2023; Chen and Zhao 2023) where the same is $\tilde{\mathcal{O}}(\epsilon^{-2})$. Note that for the latter schemes, there are no formal proofs available for the asymptotic stability and almost sure convergence (see Section 4 and Appendix of (Panda and Bhatnagar 2024) for details).
(*c*) We perform a novel asymptotic analysis of convergence of this scheme by showing that the slower timescale critic recursion remains stable and tracks a limiting differential inclusion that depends on the set of local maxima of the actor recursion corresponding to any critic update. Such an analysis under Markov noise has not been previously carried out in the context of any AC algorithm and is a generalization of the ODE based analysis of such algorithms in the presence of multiple attractors of the actor (Aubin and Frankowska 2009; Benaïm, Hofbauer, and Sorin 2005). We mention here that unlike us, most papers on finite-time analysis of RL algorithms do not prove stability and almost sure convergence of such algorithms, see Table 1. As a result, we provide stronger guarantees than such algorithms. See Appendix of (Panda and Bhatnagar 2024) for details of this analysis.
(*d*) We show the results of experiments comparing our CA algorithm with some other well-studied algorithms, on three different OpenAI Gym environments and observe that CA performs better than the other algorithms in average reward performance. In terms of training time, the CA algorithm performs uniformly better than AC requiring half the training time on two of the environments (see Section 6 and Appendix of (Panda and Bhatnagar 2024).

Notation: For two sequences $\{c_n\}$ and $\{d_n\}$, we write $c_n = \mathcal{O}(d_n)$ if there exists a constant $P > 0$ such that $\frac{|c_n|}{|d_n|} \leq P$. To further hide logarithmic factors, we use the notation $\tilde{\mathcal{O}}(\cdot)$. Without any other specification, $\|\cdot\|$ denotes the ℓ_2-norm of Euclidean vectors. $d_{TV}(M, N)$ is the total variation norm distance between two probability measures M and N, and is defined as $d_{TV}(M, N) = \frac{1}{2}\int_{\mathcal{X}} |M(dx) - N(dx)|$.

2 Related Work

We briefly review here some of the related work. In (Konda and Borkar 1999), AC algorithms were presented for the look-up table representations and the first asymptotic analysis of these algorithms was carried out. Subsequently, (Konda and Tsitsiklis 2003) presented AC algorithms with function approximation using the Q-value function and an asymptotic analysis of convergence was presented. In (Kakade 2001), a natural gradient based algorithm was presented. Subsequently, works such as (Castro and Meir 2009) and (Zhang et al. 2020) have also carried out the asymptotic analysis of AC algorithms. In (Bhatnagar et al. 2009), natural AC algorithms were presented that perform bootstrapping in both the actor and the critic recursions, and an asymptotic analysis of convergence including stability was provided. A new method for solving two-timescale optimization that achieves faster convergence was recently proposed in (Zeng and Doan 2024).

The CA algorithm was introduced in (Bhatnagar, Borkar, and Guin 2023) for the look-up table case. In this, the actor recursion is on the faster timescale compared to critic and the infinite horizon discounted cost criterion is considered. Asymptotic stability and almost sure convergence of the algorithm is shown. In our work, we present the first CA algorithm for the case of (a) function approximation and (b) the long-run average reward setting. Further, we present both – asymptotic as well as non-asymptotic convergence analyses of the proposed scheme where we observe that our algorithm gives a better upper bound on the sample complexity as opposed to AC. We also observe that our algorithm performs on par and is in fact slightly better than the two-timescale AC algorithm.

During the past few years there has been significant research activity on finite-time analysis of various algorithms in RL. A finite-time analysis of a two-timescale AC algorithm under Markovian sampling has been conducted in (Wu et al. 2022) and a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ for convergence to an ϵ-approximate stationary point of the performance function has been obtained.

Finite-time analyses of a single-timescale AC algorithm have been presented in (Olshevsky and Gharesifard 2023; Chen and Zhao 2023). In these algorithms, the actor and the critic recursions proceed on the same timescale but there are no proofs of stability and almost sure convergence of the recursions. A prime reason here is that AC algorithms are based on the policy iteration procedure whereby one ideally requires convergence of the critic in between two updates of the actor. Such guarantees can usually be obtained when there is a timescale difference between the two updates. A sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ is obtained in (Olshevsky and Ghare-

Reference	Algorithm	Sampling	Asymptotic Analysis	Sample Complexity	Critic
(Wu et al. 2022)	Two-timescale AC	Markovian	Shown in (Bhatnagar et al. 2009)	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	TD(0)
(Olshevsky and Gharesifard 2023)	Single-timescale AC	i.i.d	Not shown	$\tilde{\mathcal{O}}(\epsilon^{-2})$	TD(0)
(Chen and Zhao 2023)	Single-timescale AC	Markovian	Not shown	$\tilde{\mathcal{O}}(\epsilon^{-2})$	TD(0)
(Suttle et al. 2023)	Two-timescale MLAC	Markovian	Not Shown	$\tilde{\mathcal{O}}(\tau_{mix}^2\epsilon^{-2})$	MLMC
Our work	Two-timescale CA	Markovian	Shown	$\tilde{\mathcal{O}}(\epsilon^{-(2+\delta)})$	TD(0)

Table 1: Comparison with related works: (Olshevsky and Gharesifard 2023) uses Discounted Reward Setting while Others are for Average Reward.

sifard 2023; Chen and Zhao 2023) for single-timescale AC. While (Olshevsky and Gharesifard 2023) considers i.i.d sampling from the stationary distribution of the Markov chain in a discounted reward setting, (Chen and Zhao 2023) makes use of Markovian sampling and works with the average reward formulation. On the other hand, we obtain a sample complexity of $\mathcal{O}(\epsilon^{-(2+\delta)})$ with Markovian sampling and in the average reward setting, where $\delta > 0$ can be made arbitrarily small. In the limit when $\delta = 0$, one obtains a single-timescale AC algorithm for which asymptotic guarantees are not available. Thus, a major contribution of our work is to provide a sample complexity of our two-timescale CA scheme that is arbitrarily close to that of single-timescale AC but while providing theoretical assurances of asymptotic stability and almost sure convergence that single-timescale AC does not provide.

Non-asymptotic convergence properties of two-timescale natural AC algorithm have been studied in (Khodadadian et al. 2023) in the look-up table case where a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-6})$ has been obtained. Finite-time analysis is helpful in finding out the optimal learning rates for different updates used in the various algorithms. Amongst other recent works, (Han, Li, and Zhang 2024), (Shen and Chen 2022) and (Zhang, Zhang, and Maguluri 2021) have also provided finite-time analyses.

Table 1 shows the comparison of our work with some of these related works. In (Suttle et al. 2023), a multi-level Monte-Carlo AC algorithm is analyzed with sample complexity of $\widetilde{\mathcal{O}}(\tau_{mix}^2\epsilon^{-2})$. However, unlike us, asymptotic stability and almost sure convergence is not shown. It is also important to note that unlike many other variants (including the single-timescale AC algorithms), the two-timescale AC algorithm, as with our two-timescale CA algorithm, possesses asymptotic stability and almost sure convergence guarantees. For our algorithm, the latter properties are shown using a differential inclusions based analysis, see Appendix of (Panda and Bhatnagar 2024) for details.

3 The Framework and Algorithm

In this section, we first discuss the Markov decision process (MDP) framework. We then present our two-timescale CA algorithm where we use linear function approximation for the value function estimates.

Markov Decision Process

We consider an MDP with finite state and action spaces that is characterised by the tuple (S, A, P, r), where S denotes the state space, A is the action space, $P(s^{'}|s, a)$ is the probability of transition from state s to $s^{'}$ under action a. Further, r denotes the single-stage reward that depends on the state s and action a at a given instant. Moreover, we let $|r(s, a)| \leq U_r$, $\forall s \in S, \forall a \in A$ where $U_r > 0$ is a constant. We consider stationary randomized policies $\pi_\theta(a|s)$, $a \in A, s \in S$ parameterised by θ. Our aim is to maximise the long-run average reward (with μ_θ being the stationary distribution):

$$L(\theta) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) = E_{s\sim\mu_\theta, a\sim\pi_\theta}[r(s, a)].$$

The differential value function $V^\theta(s), s \in S$ is defined as (with s_0 being the starting state, $a_t \sim \pi_\theta(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$): $\quad V^\theta(s) = E\left[\sum_{t=0}^{\infty}(r(s_t, a_t) - L(\theta))|s_0 = s\right]$.

The differential action-value (Q-value) function is defined as

$$Q_\theta(s, a) = \mathbb{E}_\theta[\sum_{t=0}^{\infty}(r(s_t, a_t) - L(\theta))|s_0 = s, a_0 = a]$$
$$\overset{\text{(i)}}{=} r(s, a) - L(\theta) + \mathbb{E}[V^\theta(s')],$$

where the expectation in (i) is taken over $s' \sim P(\cdot|s, a)$.

The policy gradient theorem (Sutton et al. 1999; Sutton and Barto 2018) gives the following expression for $\nabla_\theta L(\theta)$:

$$\nabla_\theta L(\theta) = \mathbb{E}_{s\sim\mu_\theta, a\sim\pi_\theta}[A_\theta(s, a)\nabla_\theta \log \pi_\theta(a|s)],$$

where $A_\theta(s, a) = Q_\theta(s, a) - V^\theta(s)$ denotes the advantage function.

Function Approximation

In order to save on the computational effort needed to find exact solutions, one often uses value function approximation

Algorithm 1: Two Timescale Critic-Actor Algorithm

Input: initial average reward parameter L_0, initial actor parameter θ_0, initial critic parameter v_0, step-size α_t for actor, β_t for critic and γ_t for the average reward estimator. Draw s_0 from some initial distribution.
for $t = 0, 1, 2, \ldots$ **do**
Take the action $a_t \sim \pi_{\theta_t}(\cdot|s_t)$
Observe next state $s_{t+1} \sim P(\cdot|s_t, a_t)$ and the reward $r_t = r(s_t, a_t)$
$L_{t+1} = L_t + \gamma_t(r_t - L_t)$
$\delta_t = r_t - L_t + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t$
$v_{t+1} = \Gamma(v_t + \beta_t \delta_t \phi(s_t))$
$\theta_{t+1} = \theta_t + \alpha_t \delta_t \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)$
end for

techniques based on linear or nonlinear function approximation architectures. We use linear function approximators here for our theoretical results. Such approximators have been found to be viable for asymptotic analyses. For instance, see (Tsitsiklis and Van Roy 1999) for an asymptotic analysis of temporal difference learning algorithms and (Bhatnagar et al. 2009) for an analysis of AC algorithms when linear function approximators are used in the average cost setting. We approximate the state-value function here using a linear approximation architecture as $\widehat{V}^\theta(s; v) = \phi(s)^\top v$, where $\phi : \mathcal{S} \to \mathbb{R}^{d_1}$ is a known feature mapping and θ is the policy parameter for the considered policy.

Two-Timescale Critic-Actor Algorithm

Algorithm 1 presents the two-timescale CA algorithm involving linear function approximation for the critic recursion. All step-sizes satisfy the standard Robbins-Monro conditions. In addition, $\beta_t = o(\alpha_t)$ for $t \geq 0$ and $\gamma_t = K\alpha_t$ for some $K > 0$, $t \geq 0$. As a result of this, the average reward and actor updates are performed on the faster timescale compared to the critic updates. The projection operator $\Gamma(\cdot)$ has been used for the estimates of the critic. Here, for any $x \in \mathbb{R}^{d_1}$, $\Gamma(x)$ denotes the projection of x to a compact and convex set $C \subset \mathbb{R}^{d_1}$. For any vector $y \in C$, we have $\|y\| \leq U_v$, where $U_v > 0$ is a constant. As mentioned earlier, the single-stage reward is a function of the current state and action taken.

4 Finite-Time Analysis

We provide, in this section, the assumptions required and the main theoretical results for carrying out a non-asymptotic convergence analysis. We also state below the main results providing the optimal learning rate and sample complexity for the two-timescale CA algorithm. The detailed proofs are given in the appendix of (Panda and Bhatnagar 2024).

Assumption 4.1. The norm of each state feature is bounded by 1, i.e., $\|\phi(i)\| \leq 1$.

The above is not a restrictive assumption since the number of states $|S|$ is finite. Thus, the requirement on features can be accomplished by replacing any features $\phi(i) \in \mathbb{R}^{d_1}, i \in S$ by $\dfrac{\phi(i)}{\max_{j \in S} \phi(j)}$. This assumption is helpful in carrying out the finite time analysis of the actor and critic recursions as it helps provide suitable upper bounds for some of the terms.

Assumption 4.2. For all potential policy parameters θ, the matrix $\mathbf{A}$ defined as under is negative definite: $\mathbf{A} := \mathbb{E}_{s,a,s'}\big[\phi(s)\big(\phi(s^{'}) - \phi(s)\big)^\top\big]$, where $s \sim \mu_\theta(\cdot)$ (the stationary distribution under policy parameter θ) and $a \sim \pi_\theta(\cdot|s), s^{'} \sim P(\cdot|s, a)$. Further, let λ_θ denote the largest eigenvalue of $\mathbf{A}$. Then $-\lambda \stackrel{\triangle}{=} \sup_\theta \lambda_\theta < 0$.

Under a given policy π, Assumption 4.2 has been shown to hold in (Tsitsiklis and Van Roy 1999) in the setting of temporal difference learning under the requirements that (a) the feature vectors are linearly independent and (b) $\Phi r \neq e$, where e is the vector of all 1's. This assumption helps give the existence and uniqueness of $v^*(\theta)$ because the following equations hold: For $s \sim \mu_\theta(\cdot), a \sim \pi_\theta(\cdot|s)$,

$$\mathbf{A}v^*(\theta) + \mathbf{b} = 0, \tag{1}$$
$$\mathbf{b} := \mathbb{E}_{s,a,s'}[(r(s,a) - L(\theta))\phi(s)].$$

Assumption 4.2 helps in carrying out a finite time analysis of the critic error.

Assumption 4.3 (Uniform ergodicity)**.** Consider a Markov chain generated as per the following: $a_t \sim \pi_\theta(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)$. Then there exist $b > 0$ and $k \in (0, 1)$ such that:

$$d_{TV}\big(P(s_\tau \in \cdot|s_0 = s), \mu_\theta(\cdot)\big) \leq bk^\tau, \forall \tau \geq 0, \forall s \in S.$$

Assumption 4.3 states that the τ-step state distribution of the Markov chain under policy π_θ converges at a geometric rate to the stationary distribution μ_θ.

Assumption 4.4. There exist $L, B, K > 0$ such that for all $s, s^{'} \in S$ and $a, a^{'} \in A$,

(a) $\big\|\nabla \log \pi_\theta(a|s)\big\| \leq B, \forall \theta \in \mathbb{R}^d$,

(b) $\big\|\nabla \log \pi_{\theta_1}(a|s) - \nabla \log \pi_{\theta_2}(a^{'}|s^{'})\big\| \leq K\|\theta_1 - \theta_2\|$, $\forall \theta_1, \theta_2 \in \mathbb{R}^d$,

(c) $\big|\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)\big| \leq L\|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \mathbb{R}^d$.

Assumptions 4.4(a) and (c) are standard in the literature on policy gradient methods, see (Wu et al. 2022). Assumption 4.4(b) implies that the randomized policy is also K-smooth in the parameter θ, in addition to being Lipschitz continuous (see Assumption 4.4 (c)).

Assumption 4.5. $\forall \theta_1, \theta_2 \in \mathbb{R}^d, \forall s \in S, \exists L_\mu > 0$ such that $\|\nabla \mu_{\theta_1}(s) - \nabla \mu_{\theta_2}(s)\| \leq L_\mu \|\theta_1 - \theta_2\|$.

Assumption 4.5 implies that the stationary distribution μ_θ is L_μ-smooth as a function of θ. This assumption is required for proving smoothness of $v^*(\theta)$ and has been adopted in (Chen and Zhao 2023). We provide sufficient conditions in Theorem A.1 of (Panda and Bhatnagar 2024) for the verification of Assumption 4.5.

Assumption 4.6. $\exists L_v > 0$ such that for any $s \in S$,

$$\|V^{\theta_1}(s) - V^{\theta_2}(s)\| \leq L_v \|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \mathbb{R}^d.$$

Assumption 4.6 is needed for deriving finite time bounds while proving convergence of actor.

Let τ_t denote the mixing time of an ergodic Markov chain. So,

$$\tau_t := \min\big\{m \geq 0 \mid bk^{m-1} \leq \min\{\alpha_t, \beta_t, \gamma_t\}\big\}, \tag{2}$$

where b, k are defined as in Assumption 4.3.

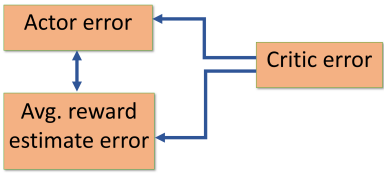

Figure 1: Dependency of errors among the actor, critic and average reward estimate.

Sample Complexity Results

We provide here the sample complexity bounds that we obtain. The proofs of these results require several detailed steps that cannot be accommodated in the limited space. Hence, we provide the complete detailed analysis in Appendix of (Panda and Bhatnagar 2024) while brief proof sketches of the main results are given here.

We consider the following step-sizes: $\alpha_t = c_\alpha/(1+t)^\nu$, $\beta_t = c_\beta/(1+t)^\sigma, \gamma_t = c_\gamma/(1+t)^\nu$ with $0 < \nu < \sigma < 1$, $2\sigma < 3\nu$, $2\sigma - \nu < 1$ and $c_\alpha, c_\beta, c_\gamma > 0$. Thus, the actor and the average reward recursions proceed here on the same timescale but which is faster than the critic recursion. Let

$$\frac{c_\alpha}{c_\gamma} < \frac{1}{2B(G+U_w)+U_wB},$$

where, $G = 2(U_r + U_v)B$, $U_w = 2B(U_v + \bar{U}_v)$ and $|V^\theta(s)| \leq \bar{U}_v, \forall \theta \in \mathbb{R}^d, \forall s \in S$, respectively.

Theorem 4.7 (Convergence of Average reward estimate)**.** *Under assumptions 4.1, 4.3, 4.4, 4.6,*

$$\sum_{k=\tau_t}^{t} \mathbb{E}[(L_k - L(\theta_k))^2] \leq \mathcal{O}(\log^2 t \cdot t^{1-\nu}) + \mathcal{O}(t^\nu)$$
$$+ 2\frac{(G+U_w)^2}{(1-\frac{c_\alpha}{c_\gamma}U_wB)^2}\frac{c_\alpha^2}{c_\gamma^2}\sum_{k=\tau_t}^{t}\mathbb{E}\|M(\theta_k, v_k)\|^2,$$

where, $L(\theta_k) = \mathbb{E}_{s\sim\mu_{\theta_k}, a\sim\pi_{\theta_k}, s'\sim P(.|s,a)}[r(s,a)]$ *and* $M(\theta_t, v_t) = E_{s_t\sim\mu_{\theta_t}, a_t\sim\pi_{\theta_t}, s_{t+1}\sim P(.|s_t,a_t)}[(r(s_t,a_t) - L(\theta_t) + \phi(s_{t+1})^\top v_t - \phi(s_t)^\top v_t)\nabla \log \pi_{\theta_t}(a_t|s_t)]$.

Proof. See Appendix of (Panda and Bhatnagar 2024). □

Theorem 4.8 (Convergence of actor)**.** *Under assumptions 4.1, 4.3, 4.4, 4.6,*

$$\frac{1}{(1+t-\tau_t)}\sum_{k=\tau_t}^{t} E\|M(\theta_k, v_k)\|^2 = \mathcal{O}(t^{\nu-1}) + \mathcal{O}(\log^2 t \cdot t^{-\nu}).$$

Proof. See Appendix of (Panda and Bhatnagar 2024). □

From Lemma 4 of (Bhatnagar et al. 2009), $M(\theta_k, v_k)$ equals the sum of the gradient of average reward and an error term that depends on the function approximator of the critic. From Theorem 4.8, convergence is to the stationary points of a function whose gradient is this sum.

Theorem 4.9 (Convergence of critic)**.** *Under assumptions 4.1, 4.2, 4.3, 4.4, 4.5, 4.6,*

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} E\|v_k - v^*(\theta_k)\|^2$$
$$= \mathcal{O}(\log^2 t \cdot t^{\sigma-2\nu}) + \mathcal{O}(t^{2\sigma-\nu-1}) + \mathcal{O}(\log^2 t \cdot t^{-3\nu+2\sigma}),$$

where $v^*(\theta_k)$ *is as defined in Equation* (1).

Proof sketch. We denote $z_t := v_t - v^*(\theta_t)$. After expanding $\|z_t\|^2$ and using Assumption 4.2, we get an upper bound for $\|z_t\|^2$ as:

$$\|z_{t+1}\|^2 \leq \|z_t\|^2 + 2\beta_t\langle z_t, \delta_t\phi(s_t) - E_{\theta_t}[\delta_t\phi(s_t)]\rangle$$
$$- 2\beta_t\lambda\|z_t\|^2 + 2\langle z_t, v^*(\theta_t) - v^*(\theta_{t+1})\rangle$$
$$+ 2\beta_t^2\delta_t^2\|\phi(s_t)\|^2 + 2\|v^*(\theta_t) - v^*(\theta_{t+1})\|^2.$$

We then rearrange the terms and take expectation of the summation from τ_t to t, to get

$$\lambda\sum_{k=\tau_t}^{t} E\|z_k\|^2 \leq \sum_{k=\tau_t}^{t}\frac{1}{2\beta_k}E[\|z_k\|^2 - \|z_{k+1}\|^2]$$
$$+ \sum_{k=\tau_t}^{t} E[\langle z_k, \delta_t\phi(s_k) - E_{\theta_k}[\delta_k\phi(s_k)]\rangle] + \sum_{k=\tau_t}^{t}\frac{1}{\beta_k}E\langle z_k,$$
$$v^*(\theta_k) - v^*(\theta_{k+1}) + (\nabla v_k^*)^T(\theta_{k+1} - \theta_k)\rangle$$
$$+ \sum_{k=\tau_t}^{t}\frac{1}{\beta_k}E\langle z_k, (\nabla v_k^*)^T(\theta_k - \theta_{k+1})\rangle$$
$$+ \sum_{k=\tau_t}^{t}\beta_k E[\delta_k^2\|\phi(s_k)\|^2]$$
$$+ \sum_{k=\tau_t}^{t}\frac{1}{\beta_k}E\|v^*(\theta_k) - v^*(\theta_{k+1})\|^2,$$

where $-\lambda = \sup_\theta \lambda_\theta$, see Assumption 4.2. After analyzing the terms on the RHS, we get the desired result. Please see the appendix of (Panda and Bhatnagar 2024) for details.

From Theorems 4.7, 4.8 and 4.9, it is clear that (as also shown in Figure 1), the critic error depends on actor error and the average reward estimate error. Moreover, actor error and average reward estimate error are dependent. Hence, Theorem 4.9 relies on the results of Theorems 4.8 and 4.7. From Theorem 4.8, we can observe that $E\|M(\theta_k, v_k)\|^2 \to 0$ as $k \to \infty$. Now as actor recursions proceed on the faster timescale as compared to the critic, the latter appears to be quasi-static to the actor, cf. Chapter 6 of (Borkar 2023). Hence, we can say that from the timescale of the actor recursion, $v_t = v, \forall t \geq 0$.Therefore the point of convergence of the actor θ_t will be $\theta(v)$ such that

$$E_{\theta(v)}[(r(s,a) - L(\theta(v)) + \phi(s')^\top v$$
$$- \phi(s)^\top v)\nabla \log \pi_{\theta(v)}(a|s)] = 0.$$

Now since the critic is on the slower timescale compared to the actor, θ_t tracks $\theta(v_t)$ at time instant t when viewed from the timescale of the critic. Moreover from Theorem 4.9, we

have $\|v_k - v^*(\theta_k)\| \to 0$ as $k \to \infty$. Hence, we can conclude that v_k converges to a point ω such that $\omega - v^*(\theta(\omega)) = 0$.

Optimizing the values of ν and σ in Theorem 4.9, we have $\nu = 0.5$ and $\sigma = 0.5 + \beta$, where $\beta > 0$ can be made arbitrarily close to zero. Hence we have the following:

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} E\|z_k\|^2 = \mathcal{O}(\log^2 t \cdot t^{(2\beta-0.5)}).$$

Therefore, in order for the mean squared error of the critic to be upper bounded by ϵ, namely,

$$\frac{1}{1+t-\tau_t}\sum_{k=\tau_t}^{t} E\|z_k\|^2 = \mathcal{O}(\log^2 T \cdot T^{(2\beta-0.5)}) \leq \epsilon,$$

we need to set $T = \tilde{\mathcal{O}}(\epsilon^{-(2+\delta)})$, where $\delta > 0$ can be made arbitrarily close to zero. For instance, $\nu = 0.5$ and $\sigma = 0.51$ gives a sample complexity $\tilde{\mathcal{O}}(\epsilon^{-2.08})$. Please see the appendix of (Panda and Bhatnagar 2024) for the detailed analysis.

Remark 4.10. The finite-time analysis of single time-scale AC cannot be easily applied here as we focus on finding the upper bound on various terms that should approach zero as time tends to infinity. We do not get such upper bounds in the sbsence of a timescale difference.

5 Asymptotic Convergence Analysis

We first note that some assumptions needed for the finite-time analysis are not required for the asymptotic convergence analysis. In particular, we do not need Assumption 4.3 on the exponential mixing of Markov noise. Our differential inclusions based asymptotic analysis that we present, unlike many other references that assume i.i.d sampling from the stationary distribution, is powerful enough to carry through under Assumptions 4.2, 4.4, and 5.1-5.2 (below). Let θ take values in a compact set $D \subset \mathbb{R}^{d_2}$.

Assumption 5.1. The Markov chain $\{s_t\}$ under any policy π^θ is ergodic for any fixed $\theta \in D$.

This assumption is routinely made for analysis of RL algorithms with Markov noise (Tsitsiklis and Van Roy 1999; Konda and Tsitsiklis 2003; Bhatnagar et al. 2009) and guarantees existence of a unique stationary distribution μ_θ for any fixed $\theta \in D$. We shall replace here the stronger requirement in Assumption 4.3 by Assumption 5.1.

Assumption 5.2. The step-size sequences $\{\alpha_t\}$, $\{\beta_t\}$ and $\{\gamma_t\}$ satisfy the following conditions:
(i) $\alpha_t, \beta_t, \gamma_t > 0$ for all t with $\gamma_t = K\alpha_t$ for some $K > 0$.
(ii) $\sum_t \alpha_t = \sum_t \beta_t = \infty$; (iii) $\sum_t (\alpha_t^2 + \beta_t^2) < \infty$.

For asymptotic convergence, we first analyze in the appendix of (Panda and Bhatnagar 2024), CA for the average reward objective, with function approximation, by incorporating a projection on the actor in addition to the critic update. Subsequently, we remove the projection on the critic update and prove the asymptotic stability and convergence of the algorithm. This algorithm is then similar to the standard AC algorithms that have been well-studied in the literature, where also one projects the actor but not the critic, cf. (Bhatnagar et al. 2009), except that now the time scales of the two recursions are reversed.

We prove the stability and convergence of our two-timescale CA algorithm by proving that the critic recursion asymptotically tracks a compact connected internally chain transitive invariant set of an associated differential inclusion (DI) (Aubin and Frankowska 2009; Benaïm, Hofbauer, and Sorin 2005). A DI-based analysis is a generalization of the ODE approach to stochastic approximation and is necessitated because we allow for multiple local maxima for the actor-recursion for any given critic update.

The critic update takes the following form, see Appendix of (Panda and Bhatnagar 2024) for details of the derivation:

$$v_{t+1} = v_t + \beta_t(y_t + \kappa_t + Y_t), \text{ where} \tag{3}$$

$y_t = \sum_s \mu_{\theta_t}(s) \sum_a \pi^{\theta_t}(s,a)(R(s,a) - L^{\theta_t} - v_t^T\phi(s)$
$+v_t^T \sum_{s'} p(s,a,s')\phi(s'))\phi(s)$, $\kappa_t = E[(R(s_t,a_t) - L^{\theta_t}$
$+v_t^T \sum_{s_{t+1}} p(s_t,a_t,s_{t+1})\phi(s_{t+1}) - v_t^T\phi(s_t))\phi(s_t)|\mathcal{F}_2(t)]$
$-y_t$, and $Y_t = -E[\delta_t\phi(s_t)|\mathcal{F}_2(t)] + \delta_t\phi(s_t)$, respectively.

Theorem 5.3 (Stability of the Critic Recursion)**.** *Under Assumptions 4.2, 4.4, 5.1 and 5.2, the recursion (3) remains stable, i.e.,* $\sup_{n\to\infty} \|v_n\| < \infty$*, w.p.1*

Proof. See Appendix of (Panda and Bhatnagar 2024). □

Consider now the following ODE associated with the faster (actor) recursion:

$$\dot{\theta} = \hat{\Gamma}_2\left(\nabla L^\theta + e^{\pi^\theta}\right), \tag{4}$$

where $\hat{\Gamma}_2(v(y)) = \lim_{0<\eta\to 0}\left(\frac{\Gamma_2(y+\eta v(y)) - y}{\eta}\right)$ and e^{π^θ} is an error term, see the appendix of (Panda and Bhatnagar 2024). Consider also the DI associated with the slower (critic) recursion (with $h(\cdot)$ as the associated set-valued map):

$$\dot{v} \in h(v), \tag{5}$$

where $h(v) = \{\sum_s \mu_\theta(s) \sum_a \pi^\theta(s,a)(R(s,a) - L^\theta$
$+v^T \sum_{s'} p(s,a,s')\phi(s') - v^T\phi(s))\phi(s)|\theta \in \bar{\theta^*}(v)\}$.

Theorem 5.4. *Suppose the ODE (4) has isolated local maxima θ^*. Correspondingly suppose $v^* \in \mathbb{R}^{d_1}$ is a limit point of the solution to the DI (5). Then under Assumptions 4.2, 4.4, 5.1 and 5.2,* $\sup_t \|v_t\| < \infty$ *and* $\sup_t \|\theta_t\| < \infty$ *w.p.1 respectively. In addition, $(v_t, \theta_t) \to (v^*, \theta^*)$ almost surely, where θ^* is a local maximum of (4) and v^* is the unique solution to the projected Bellman equation corresponding to the policy π^{θ^*}, i.e., the two together satisfy*

$$\Phi^T D^{\theta^*} \Phi v^* = \Phi^T D^{\theta^*} T_{\theta^*}(\Phi v^*). \tag{6}$$

Remark 5.5. We assume isolated local maxima for (4) in Theorem 5.4 as it helps uniquely identify the converged policy. In the absence of this assumption, one will again obtain a DI (instead of the ODE), whose limit points the algorithm will asymptotically converge to almost surely.

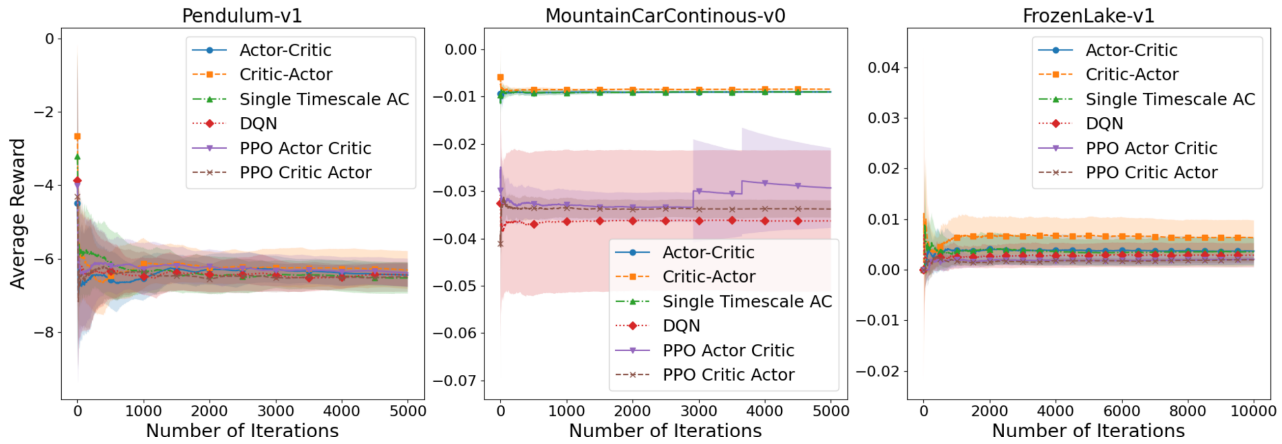

Figure 2: Comparison of Critic-Actor with few other algorithms

Environment	**Critic-Actor**	**Actor-Critic**	**DQN**	**PPO AC**	**PPO CA**	**Single Timescale AC**
Frozen Lake	0.00633 ± 0.0034	0.0036 ± 0.0027	0.0028 ± 0.0023	0.00207 ± 0.0009	0.00194 ± 0.0005	0.0035 ± 0.0028
Pendulum	-6.30 ± 0.41	-6.45 ± 0.324	-6.45 ± 0.316	-6.39 ± 0.38	-6.44 ± 0.51	-6.53 ± 0.42
Mountain Car Continuous	-0.0084 ± 0.0001	-0.009 ± 0.0002	-0.036 ± 0.014	-0.029 ± 0.0084	-0.0337 ± 0.0018	-0.009 ± 0.0002

Table 2: Comparison of Critic-Actor with different algorithms in terms of average reward

6 Experimental Results[1]

We present here the results of experiments on three different (open source) OpenAI Gym environments namely Frozen Lake, Pendulum and Mountain Car Continuous, respectively, over which we compare the performance of CA with AC as well as the Deep Q-Network (DQN) (Mnih et al. 2015) in the average reward setting, and PPO (Schulman et al. 2017). Detailed descriptions of these environments can be found at https://gymnasium.farama.org/.

While (Bhatnagar, Borkar, and Guin 2023) analyzes the asymptotic convergence of the full-state CA (FS-CA) in the discounted cost setting, for experiments, they also incorporate a setting with function approximation. For the CA and AC implementations, we have thus used their code[2] but made changes to incorporate the average reward setting. For DQN, we have used the original code from the paper and made changes to incorporate the average reward setting. For PPO, we implement two variants, namely, PPO-AC and PPO-CA, where in both algorithms, clipping has been used in the actor updates and the advantage function is estimated using the critic parameter and we have used two separate losses (the actor-loss and the critic-loss), to train the actor and the critic networks respectively. We have used the average reward setting for implementing PPO (actor and critic) unlike the base implementation that considers discounted reward.

The plots of our experiments are averaged over 10 different initial seeds after training the agent for 10,000 steps. Table 2 presents the average reward along with standard error (obtained upon convergence) for all the five algorithms in the three environments while Table 9 in the Appendix of (Panda and Bhatnagar 2024) presents their training time (in seconds). It can be seen from Table 2 that CA shows the best results in all environments, though by small margins. In terms of training time performance (Table 9 in appendix of (Panda and Bhatnagar 2024)), CA is better than AC and single-timescale AC on all three environments and in fact, it takes about half the run-time on two of the environments and is also better than the other algorithms as well except DQN. The latter has the best training time performance though it loses out on accuracy.

7 Future Work

We used a projected critic like (Wu et al. 2022; Olshevsky and Gharesifard 2023; Chen and Zhao 2023), for our non-asymptotic analysis. It would thus be of theoretical interest to derive similar bounds on the critic as we did but when projection is not used. It would also be of interest to develop potentially more efficient algorithms of the CA type, such as Natural CA, Soft CA etc., and study their theoretical convergence properties as well as empirical performance.

[1]The code for all of our experiments is available at https://github.com/prashu1306/Critic-Actor.

[2]https://github.com/gsoumyajit/Actor-Critic-Critic-Actor

Acknowledgments

The authors would like to thank the program chairs and the reviewers for their comments that helped in improving the quality of this paper. The authors were supported by the Walmart Centre for Tech Excellence, Indian Institute of Science. S. Bhatnagar was supported additionally by a J.C. Bose Fellowship, the Kotak-IISc AI/ML Centre, Indian Institute of Science, Project No. DFTM/02/3125/M/04/AIR-04 from DRDO under DIA-RCOE, and the Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science.

References

Aubin, J.; and Frankowska, H. 2009. *Differential inclusions*. Springer.

Benaïm, M.; Hofbauer, J.; and Sorin, S. 2005. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1): 328–348.

Bhatnagar, S.; Borkar, V.; and Guin, S. 2023. Actor-Critic or Critic-Actor? A Tale of Two Time Scales. *IEEE Control Systems Letters*, 7: 2671–2676.

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

Borkar, V. S. 2023. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Castro, D. D.; and Meir, R. 2009. A Convergent Online Single Time Scale Actor Critic Algorithm. arXiv:0909.2934.

Chen, X.; and Zhao, L. 2023. Finite-time analysis of single-timescale actor-critic. arXiv:2210.09921.

Han, Y.; Li, X.; and Zhang, Z. 2024. Finite-Time Decoupled Convergence in Nonlinear Two-Time-Scale Stochastic Approximation. arXiv:2401.03893.

Kakade, S. 2001. A Natural Policy Gradient. In Dietterich, T.; Becker, S.; and Ghahramani, Z., eds., *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Khodadadian, S.; Doan, T.; Romberg, J.; and Maguluri, S. 2023. Finite-Sample Analysis of Two-Time-Scale Natural Actor–Critic Algorithm. *IEEE Transactions on Automatic Control*, 68(6): 3273–3284.

Konda, V.; and Borkar, V. 1999. Actor-Critic–Type Learning Algorithms for Markov Decision Processes. *SIAM J. Control and Optimization*, 38: 94–123.

Konda, V.; and Tsitsiklis, J. 2003. OnActor-Critic Algorithms. *SIAM Journal on Control and Optimization*, 42(4): 1143–1166.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Graves, A.; Riedmiller, M.; Fidjeland, A.; and Ostrovski, G. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Olshevsky, A.; and Gharesifard, B. 2023. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2): 980–1007.

Panda, P.; and Bhatnagar, S. 2024. Two-Timescale Critic-Actor for Average Reward MDPs with Function Approximation. arXiv:2402.01371.

Puterman, M. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

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

Shen, H.; and Chen, T. 2022. A Single-timescale Analysis for Stochastic Approximation with Multiple Coupled Sequences. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 17415–17429. Curran Associates, Inc.

Suttle, W.; Bedi, A.; Patel, B.; Sadler, B.; Koppel, A.; and Manocha, D. 2023. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level Monte Carlo actor-critic. In *International Conference on Machine Learning*, 33240–33267. PMLR.

Sutton, R.; and Barto, A. 2018. *Reinforcement learning: An introduction*. MIT press.

Sutton, R.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

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

Wu, Y.; Zhang, W.; Xu, P.; and Gu, Q. 2022. A Finite Time Analysis of Two Time-Scale Actor Critic Methods, arXiv: 2005.01350.

Zeng, S.; and Doan, T. 2024. Fast two-time-scale stochastic gradient method with applications in reinforcement learning. In Agrawal, S.; and Roth, A., eds., *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, 5166–5212. PMLR.

Zhang, S.; Liu, B.; Yao, H.; and Whiteson, S. 2020. Provably Convergent Two-Timescale Off-Policy Actor-Critic with Function Approximation. arXiv:1911.04384.

Zhang, S.; Zhang, Z.; and Maguluri, S. T. 2021. Finite Sample Analysis of Average-Reward TD Learning and Q-Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 1230–1242. Curran Associates, Inc.